Synthèse 15e atelier Dialogu'IST RENATIS -13 juin 2024 " Comment garantir l'intégrité scientifique des données de recherche ? "

Sylvie Grésillaud, Fabien Borget, Magali Damoiseaux, Justine Fabre, Joanna Janik, Marie Roger-Chantin, Claire Tignolet

Synthèse 15e atelier Dialogu'IST – 13 juin 2024

« Comment garantir l'intégrité scientifique des données de recherche ? »

Groupe de travail :

Justine ANCELIN-FABRE orcid.org/0000-0001-8035-173X, Fabien BORGET orcid.org/0000-0001-7828-7046, Magali DAMOISEAUX orcid.org/0000-0002-6008-1012, Sylvie GRESILLAUD orcid.org/0000-0001-6806-4893, Joanna JANIK orcid.org/0000-0001-7587-2995, Marie ROGER-CHANTIN orcid.org/0000-0002-3384-9110, Claire TIGNOLET orcid.org/0000-0003-3405-3891

Dialogu'IST | Information Scientifique et Technique

Introduction par Fabien Borget, professeur à Aix-Marseille Université et chargé de mission science ouverte orcid.org/0000-0001-7828-7046

Parce que la recherche scientifique est une activité humaine, qui s'insère inévitablement dans un contexte sociétal, ses résultats doivent impérativement pouvoir être remis en cause, et les façons dont ils ont été obtenus, évaluées, confrontées, et reproductibles. Ces principes ont beau faire consensus dans la communauté scientifique, l'écosystème actuel de la recherche, et notamment, la façon dont recherche et chercheurs sont évalués (on peut notamment penser à l'injonction à publier, le fameux *publish or perish*), peut pousser à la méconduite. Par exemple, l'éditeur Wiley a ainsi récemment annoncé devoir rétracter plus de 11 000 articles douteux, dont une bonne partie a probablement été rédigée à l'aide de techniques d'intelligence artificielle.

Il semble plus que jamais nécessaire de pouvoir différencier les manquements volontaires à l'intégrité scientifique des méconduites involontaires. Il existe heureusement des textes de référence pour aider à cette distinction, parmi lesquels le Code de conduite européen pour l'intégrité en recherche (ALLEA, 2017-2023). Fiabilité, honnêteté, respect et responsabilité sont les quatre piliers de ce Code, et doivent également constituer les quatre piliers de toute recherche. Ces principes sous-tendent également le décret de décembre 2021 sur l'intégrité scientifique. Ce texte concerne aussi bien les individus que les institutions de recherche, qui sont tenues de fournir à leurs employés le cadre et les moyens nécessaires aux bonnes pratiques de recherche, notamment en termes d'infrastructures techniques.

Les organismes de recherche doivent aussi faire preuve de transparence : la science ouverte a en effet un lien très fort avec les pratiques d'intégrité scientifique. Dans cette optique, les données de recherche occupent une place centrale, aux côtés des résultats finaux que sont les publications. Les données doivent être préservées, mais aussi suffisamment bien documentées pour être reproduites.

Les institutions de recherche se sont emparées de ces problématiques, et commencent à décliner les recommandations internationales à leur propre échelle : elles nomment des référents à l'intégrité scientifique, s'organisent pour donner un accès FAIR aux données produites en leur sein, etc. Elles commencent également à réfléchir de concert aux évolutions à apporter au système d'évaluation de la recherche, notamment au sein de la coalition internationale CoARA (*Coalition on Advancing Research Assessment*). Mais au fond, quel est le principal enjeu de l'intégrité scientifique ? la bonne gestion et mise à disposition des données de recherche, ou le soin apporté à la publication finale ? Comment faire face aux enjeux parfois imprévisibles, qui poussent certains chercheurs à prendre des décisions dans l'urgence ?

Les différentes interventions de cet atelier dédié à la relation entre intégrité scientifique et science ouverte aborderont les aspects politiques et des retours d'expérience dans le cadre de différentes disciplines scientifiques pour répondre à ces questions.

Qualité et accessibilité des données : des pratiques de recherche responsables, garantes de l'intégrité scientifique ? Carole Chapin (Office français de l'intégrité scientifique)

L'Ofis est un département du Hcéres (Haut Conseil de l'évaluation de la recherche et de l'enseignement supérieur), qui contribue à la **définition et à la mise en œuvre d'une politique nationale de l'intégrité scientifique**. Il a également pour rôles de suivre la mise en œuvre de cette politique en collectant des données sur ce sujet, dans le but de trouver des solutions, produire des ressources et être un lieu de coopération et de coordination à l'échelle institutionnelle, nationale et européenne.

L'indissociabilité entre science ouverte et intégrité scientifique figure dans l'article D.211-3 du décret Intégrité scientifique de décembre 2023, ainsi que dans la loi pour une République Numérique de 2016, qui alimente également le code de la recherche. Le Plan national pour la science ouverte, qui n'est pas un texte de loi, précise quant à lui l'enjeu de l'accès ouvert aux publications, et préconise de faire évoluer le système d'évaluation pour ne plus avoir de pratiques contradictoires.

Depuis le décret de décembre 2023, chaque établissement public de recherche DOIT désigner un(e) référent(e) intégrité scientifique. Ce(tte) dernier(e) n'a pas pour seule mission de traiter les problèmes qui lui sont soumis : son rôle est avant tout de suggérer des pistes d'amélioration en réalisant des diagnostics, y compris dans des établissements qui ne font remonter aucune difficulté particulière. Son action est indissociable des délégués à la science ouverte.

Alors qu'on a tendance à superposer ouverture de la recherche (de ses données, en particulier) et intégrité scientifique, il faut en réalité croiser plusieurs actions pour créer de la recherche intègre, vérifiable et traçable. Il importe d'ouvrir les données et leurs métadonnées associées (dans lesquelles on peut constater des problèmes alors que les données elles-mêmes sont fiables), mais aussi les méthodes selon lesquelles elles ont été produites ou obtenues. Il importe ainsi de s'interroger sur le stockage des données, la transparence de leur mise à disposition, la citabilité de ces données (à l'aide de DOI notamment) pour en faire des travaux de recherche valorisés et valorisables au même titre que les publications finales…

Le Code de conduite européen pour l'intégrité en recherche s'articule en deux parties, l'une tournée vers les « bonnes pratiques », l'autre vers les « manquements ». Il importe toutefois de souligner qu'elles ne sont pas superposables : on ne met pas en place des bonnes pratiques uniquement pour éviter des problèmes, mais aussi dans une démarche proactive d'amélioration constante. Il est par ailleurs compliqué d'affirmer que ne pas avoir mis en place telle ou telle action est un manquement à l'intégrité scientifique.

On observe régulièrement des manquements à l'intégrité scientifique dus à la non-accessibilité des données. Depuis un règlement européen de 2014, il est obligatoire de mettre à disposition les données produites lors d'essais cliniques, notamment pour augmenter la transparence des processus, éviter la duplication inutile d'expériences, et augmenter la confiance de la société. Mais en réalité, seuls 25% des essais cliniques menés par des institutions académiques ont donné lieu à des résultats ouverts (contre 75% chez les promoteurs industriels).

En outre, s'il est important d'ouvrir des données, il est encore plus crucial d'ouvrir des données fiables : pas de données truquées, falsifiées, fabriquées par IA (cf. *paper mills*), etc. Les métadonnées ont un rôle indispensable à jouer en ce sens, et doivent-elles aussi être FAIR et intègres. Il est également fondamental d'ouvrir et documenter les résultats négatifs, et de bien le faire, car absolument toutes les données sont concernées par l'intégrité scientifique, et pas seulement celles qui vont dans le bon sens. Il est probable que les données liées aux *data paper* vont prendre de plus en plus d'importance dans les années à venir.

Il importe également de distinguer intégrité de la production scientifique (*scholarly record*) et intégrité scientifique en elle-même. On peut avoir un projet mené avec intégrité scientifique, mais dont certains éléments ne sont pas forcément intègres en eux-mêmes et doivent être corrigés. Les corrections, rétractations et errata (CRE) permettent de ne pas réutiliser des données non fiables. Ce genre de réutilisation peut être fait de bonne foi (lorsque le chercheur n'a pas connaissance des CRE), par négligence (le chercheur n'a pas vérifié l'intégrité des productions consultées), par manque d'accès à l'information, ou par « vrai » manquement à l'intégrité scientifique, lorsqu'on connaît l'existence des CRE mais tout en choisissant de les ignorer.

Mais la *reviewer fatigue* s'étend à tous les domaines. L'intégrité scientifique ayant beau être au cœur de leur métier, est-ce vraiment aux chercheurs de vérifier toutes les données et matériaux de recherche qu'ils ont sous les yeux ? Comment automatiser au maximum les vérifications pour alléger leur fardeau ? Le rôle joué par les identifiants, qui servent notamment à lier données, publications et autres matériaux de recherche, est ici indispensable. Ainsi, si par exemple une publication est rétractée, tous les jeux de données qui lui sont liés pourront être mis en doute pour peu que des DOI fassent le lien entre ces différents items.

L'intégrité scientifique modifie aujourd'hui en profondeur les codes de l'examen par les pairs, y compris dans l'évaluation des demandes de financement. Plus que jamais, il importe de penser l'écosystème de la recherche comme un tout, afin de simplifier au maximum ces questions d'intégrité en les impliquant à chaque étape du processus de recherche, et en faisant dialoguer ces étapes entre elles.

Science ouverte, intégrité scientifique et données de la recherche dans le biomédical. Gwenaël Dumont (Irset UMR Inserm 1085, membre de l'atelier rennais de la donnée ARDoISE)
orcid.org/0009-0001-4788-6342

La science ouverte à l'Inserm s'appuie sur la signature de la déclaration de Berlin (2003), le libre-accès à la connaissance, l'ouverture des publications sur HAL, le modèle de Plan de Gestion des Données (PGD) Inserm sur DMP Opidor, l'entrepôt de la donnée sur Recherche Data Gouv pour que les scientifiques puissent déposer des jeux de données. LORIER est le programme du plan stratégique 2025 de l'Inserm et dans ce cadre des affiches de sensibilisation ont été réalisées avec la constitution de groupes d'animation scientifique. L'Inserm s'est également engagé dans l'établissement d'un baromètre de la Science Ouverte (BSO).

Le HARKing est l'acronyme de Hypothesizing After the Results are Known. Ce concept a été inventé par le psychologue social Norbert Kerr qui fait référence à la pratique de recherche discutable consistant à émettre des hypothèses après que les résultats soient connus. **Les données de santé sont des données à caractère personnel particulières car considérées comme sensibles** (biomédicales, sociologiques, prestations de santé, environnement…). Si elles sont à caractère personnel, c'est qu'elles n'ont pas été anonymisées et donc il faut prendre des précautions. **Il faut des hébergeurs certifiés données de santé pour traiter ce type de données**.

L'atelier rennais de la donnée ARDoISE impacte 61 laboratoires du site rennais. **Il accompagne les bonnes pratiques sur tout le cycle de vie de la donnée**. Le PGD permet de planifier ses recherches. Pourquoi partager et ouvrir les données de santé ? Pour favoriser la réutilisation, augmenter la visibilité, participer à la transparence et accroître la confiance des citoyens, contribuer à la reproductibilité. Diffuser des résultats négatifs permet d'explorer de nouvelles hypothèses et des données à caractère unique mais aussi de répondre aux obligations légales et aux demandes des financeurs et de certaines revues. Le mot d'ordre est « **données aussi ouvertes que possible, aussi fermées que nécessaire** ».

Quand les métiers dialoguent ! Entre qualité des données et transparence de la démarche pour garantir l'intégrité scientifique des produits de la recherche en archéologie. Stéphane Renault (Coordinateur éditorial et soutien à l'édition des données à la MMSH Aix-en-Provence (CNRS - LAMPEA))

Comment la mise en application de la science ouverte pousse à la rencontre entre les métiers (approche interdisciplinaire) et comment répondre au besoin d'intégrité des données de recherche ? La Déclaration de Barcelone signée en avril 2024 par une trentaine d'institutions de recherche fait de **l'ouverture la règle par défaut pour les informations de recherche**. C'est un cadre supplémentaire pour développer de bonnes pratiques. Il existe des outils et des services d'accompagnement et de formation des communautés métiers et disciplinaires. **Les ateliers de la donnée permettent de mettre en avant les compétences locales et de les rapprocher**.

Le cycle de vie de la donnée permet aussi d'adopter de bonnes pratiques dans la vie du projet de recherche en mobilisant les personnes et les compétences. Il est nécessaire de **préparer les données en amont**. **La transmission des compétences est importante** pour que tout le monde travaille sur les données, pas uniquement les ingénieurs et que les porteurs de projet s'approprient les données.

« La pyramide de l'accès ouvert » est un projet de M. Ribary (2021) qui s'appuie à la base sur un entrepôt de projets qui va être ouvert afin de solliciter une participation et une correction. On accumule les versions sans détruire la version initiale. Ensuite des jeux de données sont constitués pour être entreposés dans un entrepôt certifié et enfin l'article scientifique va chapeauter cette pyramide. Chaque action de la recherche est nettement séparée.

Prenons l'exemple d'un groupement de recherche – le GDR SILEX (qui existe depuis 2019) – dont l'objectif est d'offrir un cadre partagé pour toute une communauté scientifique. Il y a ainsi un développement méthodologique, d'outils et de référentiels, de stratégies de publication et de référencement. L'objectif est d'avoir des outils harmonisés pour la recherche (plateforme cartographique consultable en ligne avec des points de collecte). Les données sont déposées dans un entrepôt (Nakala, CEDRE, data center AMU) et l'on réalise ensuite un data papers. Le vocabulaire est unifié afin de l'intégrer dans un thésaurus puis dans un vade-mecum qui sera publié prochainement. **La chaîne de production est transparente du début à la fin.**

A côté de la publication scientifique de synthèse (revues spécialisées académiques) **le traitement de la donnée est chronophage mais essentiel** (entrepôts de données institutionnels). Il ne faut pas oublier non plus les articles de données, techniques/méthodologiques (PGD, data/protocol/methodological papers), guides, etc. En fin de processus il y a la communication et les sites compagnons, mais attention, il faut assurer la pérennité numérique pour ces sites comme cela est demandé pour les entrepôts de données.

Il faut faire travailler ensemble des compétences pour aboutir à des stratégies communes au sein d'AMU (guichet de la donnée, cellule science ouverte, réseau d'ingénieurs de la donnée). Cet exemple montre que l'on peut aboutir à la création d'une plateforme (inter)disciplinaire avec plusieurs axes : formation, soutien, indexation et valorisation. **Toutefois, le cycle de vie est un idéal mais il est rarement atteint**. Il faut noter la nécessité d'utiliser un identifiant unique et l'importance du dialogue interdisciplinaire.

Open Science dans la physique des particules : un cas d'étude avec l'expérience ATLAS au LHC. Louie Corpe (Chaire Professeur Junior à l'Université Clermont Auvergne (UCA) / Laboratoire de Physique Clermont Auvergne (LPCA, UCA – CNRS Nucléaire & Particules)) orcid.org/0000-0003-2136-4842

L'objectif du LHC (Large Hadron Collider) est de comprendre les particules et leurs interactions. C'est une infrastructure qui produit beaucoup de données (40 millions de collisions par seconde). **On ne peut pas tout stocker donc au moment même de la collision il faut faire un choix de ce que l'on va garder**. Le trigger est au cœur de la machine, c'est le **système de déclenchement qui permet de choisir et d'orienter les données**.

Une fois les données sauvegardées, il faut les analyser, plusieurs étapes sont nécessaires avant d'arriver à l'analyse scientifique. **Chaque membre contribue au traitement global des données de la collaboration**, qui nécessite une puissance de calcul importante. Les données sont analysées à une échelle mondiale. Chacune des analyses est faite par une dizaine de personnes mais au sein d'une collaboration internationale, pionnière en termes d'open access. Tous les membres de la collaboration sont cités dans un papier car il est difficile de savoir qui a réellement fait quoi.

Mais le papier n'est en général par suffisant pour pouvoir ré-exploiter des résultats. Le **portail HEPData est ouvert et gratuit et à destination de l'ensemble de la communauté**, sur lequel les données sont mises dans un format qui peut être digeste hors de la collaboration avec pour objectif la **conservation et l'exploitation des données**. Comment passer d'un modèle à un résultat sans refaire toute l'analyse ? Sur le portail HEPData, les données sont issues des publications et des stratégies de réinterprétation. Notons l'enjeu des méta-organisations, comment cela s'organise et quelles stratégies sont mises en place ? Deux coordinateurs et des sous-groupes de travail existent pour pouvoir organiser le traitement des données. **Tout ce qui est présent dans HEPData doit être revu par les membres de la collaboration et les comités**.

Des recommandations sont effectuées par le projet ATLAS (détecteur en physique des particules installé sur le LHC) sur ce qui doit être gardé pour que le résultat puisse être réexploité. **Quels sont les dangers de manquement à l'intégrité scientifique** ? Les membres de la collaboration ATLAS sont très nombreux (jusqu'à 3500). Comment s'assurer que les personnes lisent ces publications (les auteurs) ? Et aussi comment savoir qui a fait quoi dans un papier ? Il faut également faire attention à la médiatisation du CERN qui peut influencer l'opinion.

Conclusion, Fabien Borget professeur à Aix-Marseille Université et chargé de mission science ouverte

La maîtrise de l'ensemble du cycle de vie de la donnée est un jalon essentiel pour garantir l'intégrité scientifique, en lien avec la mise à disposition des données et leur FAIRisation. On cartographie et l'on centralise des compétences autour de la donnée. On constate une multitude de dépôts autour des données, on a besoin d'identifier ces ensembles et d'assurer leur interopérabilité. L'attribution d'identifiants pérennes doit permettre de mettre en connexion les chercheurs et les laboratoires mais aussi de connecter les données, de relier les protocoles, les méthodes et la méthodologie. C'est cette transparence qui sera garante de l'intégrité scientifique. Il faut noter l'importance de la réforme de l'évaluation de la recherche (par exemple avec 3500 signataires d'un papier ; les 11 000 papiers rétractés par Wiley). L'ouverture des publications et des données sont garantes de l'intégrité scientifique.