Le rôle des dépôts dans le maintien de l’intégrité de la recherche

Traduction en français de The role of repositories in maintaining research integrity
https://zenodo.org/records/20718468

CCSD – June 2026

Pression sur la production/données scientifique

La science repose sur un socle de confiance. Chercheurs, institutions, éditeurs et grand public s’appuient sur l’hypothèse que les résultats publiés sont le fruit de recherches honnêtes, rigoureuses et transparentes. Pourtant, ce socle est de plus en plus menacé. Dans toutes les disciplines et institutions du monde entier, les problèmes d'intégrité de la recherche prennent de l'ampleur et se complexifient, soulevant des questions quant à la fiabilité des données scientifiques et à la crédibilité de la recherche elle-même.

Les violations les plus graves de l’intégrité scientifique sont les inventions et les falsifications, qui consistent à fabriquer ou manipuler des données pour étayer des conclusions souhaitées (Rubin, 2022). Bien que ces cas aient historiquement été considérés comme rares (National Academies of Sciences, Engineering, and Medicine, 2017), leurs conséquences peuvent être dévastatrices. En médecine et en santé publique, des résultats falsifiés ont conduit à des traitements dangereux, à une mauvaise allocation des ressources et à une perte durable de la confiance du public (Mousavi et Abdollahi, 2020 ; Motta et Stecula, 2021). Au-delà de ces cas isolés, une industrie en pleine expansion s’est développée autour de la production de faux articles. Les « usines à articles » (*paper mills*), qui vendent aux chercheurs des manuscrits entièrement fabriqués, représentent désormais une menace systémique pour la production scientifique (Parker et al., 2024 ; Richardson et al., 2025). Dans certains jeux de données ouverts, les articles produits en masse ont été estimés jusqu’à dix fois plus nombreux que les articles de recherche authentiques (Barnett et Spick, 2026).

Les éditeurs traditionnels, submergés par les articles frauduleux, consacrent désormais des ressources considérables à leur détection, mais beaucoup passent encore inaperçus lors du contrôle éditorial (Bik, 2026). L’émergence de l’IA agentique (MIT Sloan School of Management, 2026) — des systèmes capables de générer des manuscrits complets sans question de recherche impulsée par l’humain — aggrave encore la situation, épuisant les ressources disponibles pour filtrer ces travaux.

Cependant, la fraude délibérée n’est pas la seule menace pesant sur l’intégrité scientifique. Même les recherches menées de bonne foi peuvent saper la confiance dans la science lorsque leurs résultats ne peuvent pas être reproduits indépendamment, un phénomène désormais largement connu sous le nom de « crise de la reproductibilité » (Caffrey, 2024). En 2015, par exemple, l’Open Science Collaboration a tenté de reproduire 100 études publiées en psychologie et a constaté que moins de la moitié produisaient des résultats cohérents avec les conclusions originales. Depuis, des préoccupations similaires ont émergé en médecine

(Errington et al., 2021) et en économie (Camerer et al., 2016). L'utilisation croissante de l'IA et de l'apprentissage automatique dans le processus de recherche ajoute une complexité supplémentaire. Comme le souligne le rapport de la Royal Society sur la science à l'ère de l'IA (2024), la reproduction des recherches basées sur l'IA nécessite non seulement de reproduire les méthodes, le code, les données et les conditions environnementales dans lesquelles l'expérience a été menée, mais aussi une transparence sur les données d'entraînement utilisées pour développer l'outil, qui peuvent introduire des biais de manière involontaire.

Le problème pourrait également s'auto-alimenter, dans la mesure où l'IA accélère la vitesse à laquelle des contenus de mauvaise qualité et inventés de toutes pièces s'intègrent dans le corpus scientifique. Les grands modèles linguistiques étant continuellement entraînés sur de nouveaux contenus, les informations inventées risquent d'être absorbées et amplifiées à chaque cycle (COAR, 2026). Avec le temps, la frontière entre les connaissances authentiques et celles qui sont inventées risque de devenir de plus en plus difficile à discerner, non seulement pour le grand public, mais aussi pour les chercheurs eux-mêmes.

Prévenir une crise

De plus en plus, les technologies qui permettent la fraude sont également mises en œuvre pour la contrer (Candal-Pedreira & Ruano-Ravina, 2026). Les outils d'IA sont désormais capables d'analyser des millions d'articles à la recherche de signatures textuelles associées à l'activité des « usines à articles » (Cabanac et al., 2022), de signaler les manipulations d'images avec une précision qui surpasse celle de l'examen humain (Chaturvedi et al., 2025) et d'identifier des incohérences statistiques (Nuijten & Polanin, 2020) qui, sans cela, passeraient inaperçues lors de l'évaluation par les pairs. Mais la détection assistée par l'IA a ses limites. La fraude et les dispositifs de détection sont engagés dans une course aux armements permanente, et un système qui repose principalement sur la détection a posteriori des manquements sera toujours à la traîne.

Une réponse plus durable se situe en amont et s'articule autour de trois axes complémentaires : des mécanismes d'incitation qui privilégient la rigueur plutôt que la quantité, des pratiques de science ouverte qui rendent le processus de recherche transparent et vérifiable, et des infrastructures qui garantissent la traçabilité et favorisent la responsabilité à long terme.

La dimension incitative

Les pratiques de science ouverte nécessaires pour renforcer l'intégrité de la recherche ne deviendront la norme que si la communauté scientifique modifie la manière dont les chercheurs sont évalués. La culture du « publier ou périr » reste l'un des principaux moteurs structurels des problèmes d'intégrité. En récompensant la quantité plutôt que la qualité, les chercheurs sont incités à maximiser leur production au détriment de la rigueur (Edwards et Roy, 2017). Lorsque l'avancement de carrière dépend du nombre de publications, les chercheurs pratiquent le « salami slicing » (découpage d'une seule étude en plusieurs articles plus petits) pour maximiser leurs résultats. Et lorsque les revues privilégient les résultats positifs, les résultats nuls ou négatifs restent non publiés, créant un biais systémique dans la base de preuves sur laquelle d'autres s'appuient. Une enquête internationale menée auprès

de plus de 3 000 chercheurs dans 45 pays a révélé que la plupart des chercheurs connaissaient les pratiques de recherche ouverte, mais que le problème résidait davantage dans les structures incitatives que dans un manque de sensibilisation (Pennington et al., 2026). L'étude a montré que la faible adoption de la science ouverte était due aux structures incitatives, au financement et au besoin de reconnaissance. Tant que ces incitations ne changeront pas, les chercheurs continueront à privilégier les résultats publiables plutôt que des pratiques de recherche robustes et transparentes.

Des efforts significatifs pour réformer l'évaluation de la recherche sont déjà en cours, notamment en Europe. La Coalition for Advancing Research Assessment (CoARA), lancée en 2022, a réuni des centaines d'universités, de financeurs et d'organisations de recherche déterminés à faire évoluer les systèmes d'évaluation pour s'éloigner des indicateurs bibliométriques réducteurs et s'orienter vers une évaluation qualitative et multidimensionnelle. Avec un soutien croissant des gouvernements et des institutions, CoARA représente l'une des initiatives structurelles les plus significatives actuellement en cours. D'autres pays, comme la Chine, modifient également leur manière d'évaluer la recherche. La Chine s'éloigne progressivement d'un recours depuis plusieurs décennies aux indicateurs quantitatifs pour s'orienter vers un ensemble plus diversifié de critères mettant l'accent sur la qualité, l'innovation et l'impact sociétal (Shu et al., 2022 ; Zhao et al., 2026). À mesure que d'autres pays suivent cet exemple, comme les Pays-Bas (Rushforth, 2025), le Royaume-Uni (Research on Research Institute, s.d.), et d'autres, on peut s'attendre à une transition progressive mais significative dans les pratiques de publication au sein de la communauté de recherche dans les années à venir. Cependant, l'impact dépendra finalement de la manière dont ces principes seront adoptés en pratique par les gouvernements et les universités.

La science ouverte comme réponse structurelle

Combinée à la réforme de l'évaluation de la recherche, la science ouverte constitue l'approche la plus efficace pour relever les défis liés à l'intégrité, car elle permet une meilleure vérification des données, du code et des méthodologies sous-jacentes aux résultats publiés. En ce sens, la transparence est plus qu'un vague principe scientifique, mais un mécanisme concret pour créer un écosystème de communication savante digne de confiance. Lorsque les produits de la recherche sont ouvertement accessibles, les erreurs et les incohérences ont plus de chances d'être identifiées et signalées par la communauté dans son ensemble. La reproductibilité devient véritablement possible car les matériaux sous-jacents sont accessibles plutôt que verrouillés sur des disques durs et disponibles uniquement sur demande. Cette ouverture crée également une trace durable : les ensembles de données, les codes et les prépublications hébergés dans des dépôts subsistent au-delà de la durée de vie de tout projet individuel ou institution, permettant ainsi aux résultats d'être réexaminés, ré-analysés, remis en question ou développés au fil du temps. Bien qu'il existe des cas d'exploitation abusive de données ouvertes (Spick et al., 2026), ce problème peut être mieux géré par un examen plus attentif des articles et des données plutôt que par un recul de l'ouverture.

La transparence des données, du code et des méthodes répond à une dimension du défi que représente l'intégrité, mais les processus utilisés pour évaluer la recherche nécessitent également une plus grande ouverture. La relecture par les pairs, pierre angulaire du contrôle qualité dans l'édition académique, pourrait contribuer de manière significative au

renforcement de l'intégrité de la recherche si elle était rendue ouvertement accessible. À son niveau le plus fondamental, rendre le processus de relecture visible en publiant le contenu des évaluations et l'identité des relecteurs crée une responsabilité bien plus grande. Les relecteurs qui savent que leurs évaluations seront lues par d'autres pourraient s'engager plus attentivement avec un manuscrit et être moins susceptibles d'approuver un travail sans examen approfondi (Walsh et al., 2000). De plus, la publication des rapports de relecture aux côtés des articles donne aux lecteurs une vision plus complète de la manière dont un article a été évalué, des réserves qui ont été émises et de la manière dont les auteurs y ont répondu, plutôt que de présenter l'acceptation comme une validation simple et opaque. L'ouverture de l'évaluation par les pairs pourrait aussi jouer un rôle préventif : sachant que les critiques formulées à l'égard de leurs travaux seront rendues publiques et que leurs auteurs seront identifiés, les chercheurs pourraient être davantage incités à s'assurer que leurs travaux résistent à un examen public.

Le modèle émergent « Publish, Review, Curate » (PRC), dans lequel les manuscrits sont publiés sous forme de prépublications, évalués en toute transparence, puis sélectionnés, représente l'une des initiatives les plus prometteuses visant à repenser l'édition scientifique dans ce sens. Le modèle PRC – qui gagne du terrain au sein de la communauté scientifique – est un terme générique désignant un ensemble varié d'approches d'évaluation ouverte.

Les preuves démontrant que ces pratiques peuvent faire la différence sont convaincantes. Une analyse interdisciplinaire a révélé que la proportion d'articles faisant état de résultats positifs corroborant l'hypothèse testée avait augmenté au fil du temps, pour atteindre environ 86 % en 2007, certains domaines affichant même des taux plus élevés (Fanelli, 2012). Plus d'une décennie plus tard, le même constat demeure. Une étude réalisée en 2021 comparant les recherches en psychologie menées de façon traditionnelle et celles ayant fait l'objet d'un enregistrement préalable, selon le modèle des « Registered Reports », a révélé que 96 % des articles traditionnels faisaient état de résultats positifs, contre seulement 44 % des études enregistrées au préalable (Scheel et al., 2021). Cela suggère que lorsque le processus de recherche est transparent dès le départ, les distorsions introduites par le biais de la publication et la communication sélective des résultats sont considérablement réduites.

Les dépôts peuvent jouer un rôle clé dans le soutien à l'intégrité de la production scientifique

L'intégrité de la recherche ne se limite pas au comportement des individus ou des institutions. Elle dépend également de l'infrastructure qui conserve, relie et rend accessibles les résultats de la recherche. Lorsque ces fonctions font défaut ou sont fragmentées, les données elles-mêmes perdent leur fiabilité, quelle que soit la rigueur avec laquelle les études individuelles ont été menées. Les dépôts jouent un rôle de plus en plus important dans ce contexte. Ils combinent les éléments structurels nécessaires au maintien de l'intégrité des données scientifiques au fil du temps : métadonnées de provenance, identification pérenne, gestion des versions, vérification des contributeurs et préservation à long terme des ressources. De plus, comme ils sont généralement rattachés à une université, à un institut de recherche ou à une communauté disciplinaire, leur gouvernance est davantage alignée sur la mission de recherche que sur des intérêts commerciaux. Bien que les chercheurs semblent encore préférer les services disciplinaires ou ceux proposés par des prestataires commerciaux, les

avantages liés à l'utilisation d'un dépôt bénéficiant d'un soutien institutionnel et d'une gouvernance communautaire deviendront probablement plus évidents dans le contexte actuel.

Provenance des données

La provenance des données est au cœur de l'intégrité de la recherche : il s'agit de l'ensemble des informations qui documentent l'origine d'une ressource, la manière dont elle a été créée et les transformations qu'elle a subies au fil du temps. Des métadonnées structurées et détaillées, associées à un accès pérenne aux ressources, permettent de distinguer les travaux scientifiques authentiques des travaux fabriqués de toutes pièces ou manipulés. Les identifiants pérennes garantissent que les citations restent accessibles et traçables, même lorsque les plateformes évoluent, ferment ou changent de mains. La provenance est renforcée lorsque les ressources associées sont directement reliées entre elles, par exemple les publications aux ensembles de données ou le code aux sources de financement, ce qui permet à d'autres de vérifier, interpréter et réutiliser ces éléments en toute confiance. La vérification des contributeurs par le biais de métadonnées structurées, y compris les identifiants ORCID et les déclarations de rôles CRediT, ajoute également de la granularité à l'enregistrement de la provenance en rendant les identités des contributeurs et les rôles déclarés plus explicites. Cependant, il y a eu des exemples de mauvaise utilisation de ces pratiques, comme des comptes fantômes sur ORCID (Teixeira da Silva, 2021).

Les conséquences d'une provenance des données insuffisante peuvent être considérables. Un preprint récent de Gibson et al. a identifié deux jeux de données largement utilisés qui ne disposaient pas d'une provenance vérifiable et présentaient des signes de fabrication. Malgré cela, ces données avaient déjà servi à développer 124 modèles de prédiction clinique publiés, dont trois ont été appliqués en pratique clinique et un a été cité dans un brevet relatif à un dispositif médical (Gibson et al., 2026). Alors que la falsification des ressources scientifiques se généralise, ce type de métadonnées structurées et lisibles par machine sur la provenance devient un mécanisme de protection essentiel.

Gestion des versions

La gestion des versions est une dimension fondamentale mais souvent négligée d'une bonne gestion de dépôt. La recherche est rarement un produit statique et ponctuel : les jeux de données sont mis à jour, le code est révisé, les preprints sont affinés et mis à jour en fonction des retours, les rapports sont modifiés à la lumière de nouvelles découvertes. En l'absence d'informations claires sur la gestion des versions, il devient impossible pour les utilisateurs de savoir à quelle version d'une ressource ils accèdent, s'il existe une version plus récente et en quoi le contenu a pu évoluer d'une version à l'autre. Cette ambiguïté comporte des risques réels. Un chercheur qui cite ou s'appuie sur une version obsolète d'un jeu de données, d'un projet d'article ou d'un modèle, sans savoir qu'une version plus récente existe, pourrait propager des erreurs dans des travaux ultérieurs.

Conserver des informations claires sur les versions renforce également l'intégrité de la recherche de manière plus générale, en créant une piste d'audit transparente qui permet de retracer la provenance et l'évolution d'une ressource au fil du temps (Klump et al., 2021). Cela

revêt une importance particulière dans la recherche computationnelle, où la reproductibilité des résultats dépend non seulement de l'accès aux données et au code, mais aussi de la connaissance précise de la version de chacun d'entre eux qui a été utilisée dans une analyse donnée.(Stodden et al., 2016).

Gestion des corrections et rétractations

Alors que le suivi des versions concerne les révisions apportées par l'auteur lui-même, les corrections et rétractations reflètent l'état d'une ressource. Cependant, une chaîne de correction rompue représente un autre risque pour l'intégrité scientifique. Lorsqu'un éditeur rétracte un article, ce même article peut continuer à circuler sans avoir été corrigé dans un dépôt, à l'insu des lecteurs qui n'ont aucun moyen de savoir que son statut a changé.

Les dépôts n'ont traditionnellement pas fait preuve d'une grande proactivité dans le suivi des rétractations, même si les récentes évolutions liées à Retraction Watch, associées à des améliorations au niveau des infrastructures, ont rendu cette tâche plus facile à gérer. En septembre 2023, Crossref a acquis la base de données Retraction Watch et, en janvier 2025, a intégré ces données à son API REST, permettant ainsi aux systèmes de vérifier si un DOI a été signalé pour rétractation (Crossref, 2025). Cette avancée technique est complétée, au niveau politique, par les bonnes pratiques recommandées par la National Information Standards Organisation (2024) dans son document intitulé « Communication of Retractions, Removals, and Expressions of Concern » (CREC), qui établit un cadre normalisé régissant la manière dont les informations relatives aux rétractations doivent circuler d'une plateforme à l'autre.

Pour les dépôts qui contiennent des articles avec des DOI d'éditeurs, des requêtes API régulières pourraient permettre de vérifier automatiquement leurs statuts afin de mettre à jour les enregistrements sans intervention manuelle (Crossref, 2025). L'outil « Article Status Monitor » d'Europe PMC (*Europe PMC's Article Status Monitor*) démontre la faisabilité de cette approche en acceptant des identifiants de publication par lots et en renvoyant des mises à jour de statut — y compris les rétractations, les retraits et les nouvelles versions —, mettant ainsi en évidence des changements qui, sans cela, pourraient passer inaperçus (Rosonovski et al., 2023 ; Levchenko et al., 2024). Comme de nombreux dépôts n'ont pas encore réfléchi à la manière de gérer les rétractations, COAR va mettre en place un groupe de travail chargé d'élaborer des recommandations pour la gestion des rétractations et des corrections dans les dépôts.

Préservation

La traçabilité de la provenance, des versions et des corrections repose sur la condition préalable que les ressources elles-mêmes restent accessibles au fil du temps ; de ce fait, la préservation numérique est une condition essentielle à l'intégrité de la recherche. Un document scientifique auquel on ne peut accéder, qu'on ne peut vérifier ou enrichir de manière fiable au fil du temps est un document auquel on ne peut se fier – et c'est précisément cette accessibilité et cette authenticité à long terme que la préservation numérique vise à garantir. Les dépôts sont parmi les réponses institutionnelles les plus efficaces à ce défi. Cependant, la préservation à long terme est coûteuse et nécessite un engagement complet et une

stratégie de la part de l'institution où le dépôt est hébergé. Pour cette raison, de nombreux dépôts ont adopté certaines pratiques qui soutiennent la préservation, mais à ce jour, n'ont pas été en mesure de mettre en œuvre des stratégies de préservation complètes.

À mesure que les dépôts assument la responsabilité de la diffusion numérique au service de l'intégrité de la recherche, ils devront consacrer plus de ressources à ces activités. Il est essentiel de noter que, la plupart des dépôts étant gérés par des universités, des instituts de recherche ou des organisations à but non lucratif, leurs engagements en matière de préservation ne dépendent pas de la viabilité commerciale ou ne sont pas exposés aux perturbations liées aux rachats et aux fermetures de plateformes qui ont affecté de nombreux services scientifiques privés.

Conclusion

L'intégrité de la science n'est pas une préoccupation marginale. C'est la condition sur laquelle tout le reste repose : la transposition de la recherche en politiques publiques, la mise au point de traitements accessibles aux patients et la confiance du public qui sous-tend l'ensemble de cette entreprise. Les défis évoqués dans cet article, allant de la fabrication et de la falsification de données aux articles générés par l'IA en passant par les méthodes non reproductibles, ne constituent pas des défaillances isolées de chercheurs agissant de mauvaise foi. Il s'agit, en grande partie, des conséquences prévisibles de systèmes qui ont trop longtemps récompensé les mauvaises pratiques.

Les dépôts sont au cœur de la réponse à ces défis. En tant qu'infrastructures pérennes, non commerciales et gérées par la communauté, ils sont idéalement placés pour répondre aux exigences actuelles : des métadonnées de provenance structurées qui rendent la manipulation plus difficile à dissimuler, une gestion des versions qui préserve la traçabilité des recherches en évolution, un système de signalement des corrections et des rétractations qui comble les lacunes des processus réservés aux éditeurs, et des liens vers des services d'évaluation par les pairs ouverts qui rendent l'évaluation de la recherche visible et responsable. Aucune de ces fonctions n'est possible sans un investissement soutenu et un engagement institutionnel fort. Or, chacune d'entre elles est essentielle.

Pour aller de l'avant, il faut agir simultanément sur plusieurs fronts : réforme des mécanismes d'incitation, pratiques de science ouverte et infrastructures robustes. La science ouverte permet l'examen critique, mais seulement si l'infrastructure garantit la traçabilité des travaux à long terme et si les systèmes d'incitation récompensent les chercheurs qui la mettent en œuvre. Les progrès dans chaque domaine renforcent les autres et, ensemble, ils convergent vers un écosystème de recherche où l'excellence scientifique et la reconnaissance de la recherche ne sont plus incompatibles.

Remerciements

Ce rapport a été rédigé par Diana Furcila, consultante COAR, et Kathleen Shearer, directrice générale de COAR, puis examiné par le conseil d'administration de COAR.