La valeur publique inexploitée des dépôts en libre accès

Traduction en français de : The Untapped Public Value of Open Access Repositories
https://zenodo.org/records/20717974

Introduction

Après de nombreuses années d'investissements soutenus, il existe désormais des milliers de dépôts en libre accès dans le monde entier, qui collectent et préservent de précieux résultats de recherche. Une analyse portant sur environ la moitié des archives répertoriées dans le répertoire international des dépôts de COAR a révélé que 3 489 dépôts, interrogés via le protocole OAI-PMH, contenaient près de 340 millions d'enregistrements couvrant une grande variété de productions scientifiques : articles, jeux de données, codes et rapports. Ces archives sont conçues pour être ouvertes et ont été développées afin de fournir un accès sans barrière aux résultats de la recherche, aussi bien pour les humains que pour les machines. Ensemble, elles constituent un pilier essentiel de l'infrastructure mondiale de la recherche, permettant aux chercheurs et au grand public d'accéder aux connaissances collectives produites par la recherche scientifique.

Au cours des deux dernières décennies, les arguments en faveur des dépôts en libre accès se sont principalement concentrés sur les avantages pour la communauté scientifique, tels que l'accélération des découvertes, le développement de carrière et l'augmentation de la visibilité académique (Romano & Boggio, 2024). Cette approche a contribué à l'adoption de nombreuses déclarations et politiques réclamant que les recherches financées par des fonds publics soient rendues librement accessibles, comme la Déclaration de Budapest (2002), la Recommandation de l'UNESCO sur la science ouverte (2021), le Plan S (cOAlition S, 2018), et la politique de l'Office of Science and Technology Policy (OSTP) (2022). Cependant, l'un des avantages les plus importants de l'accès ouvert — l'accès du public aux résultats de la recherche — a souvent été négligé.

Les connaissances scientifiques au-delà du monde académique

L'idée que la recherche scientifique doit contribuer au progrès sociétal et industriel n'est pas nouvelle, mais elle a traditionnellement été mise en œuvre par des activités telles que la commercialisation ou la mobilisation des connaissances, des processus visant à transformer les résultats de la recherche en produits, politiques et pratiques exploitables.

Cependant, un ensemble important et croissant de données suggère que lorsque les résultats de la recherche sont mis en libre accès directement à la disposition du public, ils sont utilisés de multiples façons par des communautés très diverses. Par exemple, la thèse d'Alperin

(2015) a révélé une utilisation intensive des portails de revues en accès ouvert d'Amérique latine (SciELO et RedALyC) par des étudiants, des non-académiques et des utilisateurs du grand public. Une analyse de la plateforme finlandaise Journal.fi a montré que seulement 36 % des utilisateurs étaient des chercheurs, les autres étant des étudiants, des citoyens, des experts externes et des enseignants (Pölonen et al., 2021). Des témoignages isolés concernant l'utilisation des dépôts ouverts montrent des modèles similaires à ceux observés avec les revues en accès ouvert. Bien que la plupart des dépôts ne collectent pas activement d'informations sur leurs utilisateurs, ceux qui le font rapportent une grande diversité d'audiences et de motivations pour accéder aux contenus.

Le dépôt DASH de Harvard, par exemple, a lancé en 2012 une rubrique intitulée « DASH Stories », invitant les lecteurs à partager la manière dont ils utilisent les ressources en libre accès disponibles dans l'archive ou en tirent profit. Les retours sont éloquents : plus de 9 000 contributions émanant de médecins, de patients, d'enseignants, d'entrepreneurs et de citoyens du monde entier offrent un portrait riche de l'engagement du public envers les ressources en libre accès. Les bibliothèques du MIT, le Digital Repository of Ireland et l'université d'Édimbourg, qui ont aussi recueilli des témoignages d'utilisateurs, ont rassemblé des témoignages similaires, soulignant la valeur considérable, mais souvent sous-estimée, des résultats de recherche lorsqu'ils sont mis à disposition en libre accès dans un dépôt.

Les exemples suivants illustrent comment le libre accès peut donner plus d'autonomie aux individus, tant dans leur vie professionnelle que personnelle :

- Un utilisateur de *Harvard DASH Stories* :

  « Je suis retourné en Inde en 2005 pour travailler avec les habitants des bidonvilles de Mumbai et les populations autochtones du district de Palghar, en collaborant avec des jeunes de la communauté sur des actions fondées sur la recherche, le développement communautaire et le renforcement de la démocratie locale. Ce type de littérature éducative nous aidera certainement à comprendre les schémas comportementaux des populations défavorisées et à élaborer des stratégies plus adaptées à leurs besoins, apportant ainsi des changements positifs et significatifs dans leur vie. Alors, un grand MERCI de rendre cette ressource accessible à ceux d'entre nous qui travaillent dans le Sud global. Dans la plupart des cas, nous ne pouvons pas nous permettre des sites comme JSTOR. MERCI ! »

- Un journaliste utilisateur de l'archive *DIGITAL.CSIC* :

  « Dans notre cas, le contenu accessible ouvertement via le portail DIGITAL.CSIC a servi de source pour documenter certains des articles publiés sur notre site web. Pouvoir trouver les données nécessaires dans des études menées par des professionnels de leur domaine et mises à disposition en ligne, ouvertement et gratuitement, est, je crois, l'une des manières les plus importantes — et pourquoi ne pas le dire, fondamentales — de sensibiliser aujourd'hui à un large éventail de

sujets, ce qui nous a été d'une grande aide. »

Le rôle des dépôts dans le maintien de la confiance du public

Parallèlement, la méfiance envers les affirmations scientifiques ne cesse de croître au sein de certaines communautés, alors que les connaissances sont de plus en plus médiatisées par les réseaux sociaux et d'autres plateformes numériques, où elles peuvent être mal interprétées ou manipulées à des fins politiques ou commerciales (Gallup, 2024 ; Edelman, 2024). Le public rencontre des difficultés réelles à distinguer les informations factuelles de la désinformation, d'autant plus que cette dernière adopte souvent les mêmes conventions que les contenus légitimes, en particulier dans un environnement dominé par l'IA.

Malgré ce scepticisme, les scientifiques jouissent encore d'un niveau de confiance relativement élevé. Une étude de Cologna et al. (2025), portant sur plus de 71 000 répondants dans 68 pays, a révélé une confiance considérable dans les scientifiques, avec 83 % des répondants estimant que les chercheurs devraient communiquer davantage avec le public. De même, une enquête du Pew Research Center (2024) indique que les trois quarts des Américains font confiance aux scientifiques, les classant bien devant les journalistes et les élus (Tyson & Kennedy, 2024).

Dans ce contexte, les dépôts, qui fournissent un accès ouvert aux résultats de la recherche produits par les chercheurs eux-mêmes, sont bien positionnés pour jouer un rôle clé dans le maintien et le renforcement de la confiance du public. Leur crédibilité est renforcée par le fait qu'ils sont généralement gérés par des institutions académiques ou à but non lucratif, qu'ils sont rarement axés sur le profit et qu'ils fonctionnent généralement selon des structures de gouvernance transparentes (L'Hours et al., 2019). Les dépôts liés à des institutions et dotés d'un périmètre défini apportent une couche supplémentaire de légitimité, car ils n'acceptent de contributions que de la part de leurs enseignants et étudiants, réduisant ainsi le risque de soumissions frauduleuses. Cette combinaison de gestion institutionnelle et d'accès ouvert aux sources primaires établit une chaîne de traçabilité qui relie les productions scientifiques à des organisations responsables, faisant des dépôts des sources d'information hautement fiables. De plus, ils représentent une extension naturelle de la mission citoyenne des universités, en offrant à leurs communautés un accès direct à leur production de recherche (Moore et al., 2020).

Fonctionnalités destinées au grand public

Malgré leur valeur indéniable pour le public, les dépôts ouverts n'ont pas été conçus à l'origine pour un public non spécialisé, mais plutôt pour les communautés de recherche, et n'ont donc pas été optimisés pour l'utilisateur lambda. Leur utilité pourrait être considérablement améliorée grâce à l'adoption de fonctionnalités plus adaptées au grand public, telles que des résumés en langage simple, des indications plus claires sur la provenance, et des connexions plus systématiques et fluides avec les services intermédiaires (moteurs de recherche, chatbots, etc.) largement utilisés par le public pour rechercher des informations.

Les résumés en langage simple sont une méthode bien établie pour améliorer la compréhension des informations de recherche par les non-experts, mais ils ne sont pas encore largement utilisés par les dépôts. Ces résumés communiquent généralement l'importance d'un article de recherche scientifique dans un langage exempt de jargon. Ils sont déjà produits par de nombreux éditeurs en biomédecine, qui ont constaté leur efficacité et leur utilisation généralisée. Une étude sur les *digests* en langage simple de la revue *eLife* a montré qu'ils étaient très appréciés des chercheurs pour obtenir des perspectives interdisciplinaires (King et al., 2017). Fait remarquable, les principaux utilisateurs étaient des chercheurs travaillant en dehors de leur domaine de spécialité. bioRxiv, un serveur de preprint dans les sciences biomédicales, a lancé un projet pilote en 2023 visant à utiliser les modèles de langage (LLM) pour améliorer l'accessibilité des contenus. Chaque preprint était accompagné de trois résumés générés par IA, chacun destiné à un type de lecteur différent :

1. Une personne sans formation scientifique (ou très limitée).
2. Un scientifique expert dans un autre domaine.
3. Une personne dont l'expertise est équivalente à celle de l'auteur.

En 2024, le projet pilote a été étendu, avec des résumés générés par IA améliorés et la possibilité pour les auteurs de les modifier. Compte tenu des progrès rapides des outils d'IA pour la synthèse d'informations, les résumés en langage simple (sous forme de texte ou de vidéo) pourraient représenter une amélioration très utile pour les archives à court terme et améliorer considérablement leur valeur pour les lecteurs non experts (Lavoie, 2025).

Les indicateurs de provenance et de crédibilité sont un autre aspect important pour garantir la confiance du public. Le rôle du dépôt en tant que point de référence fiable pour les utilisateurs est renforcé lorsque certaines conditions au niveau de l'infrastructure sont remplies. Des métadonnées bien structurées et complètes, un engagement en faveur de la conservation à long terme, des liens persistants via des identifiants pérennes (PIDs), des connexions entre ressources connexes, et des formats et API adaptés aux machines sont tous indispensables pour fournir aux utilisateurs des informations contextuelles sur la ressource et sa fiabilité.

Sans ces éléments, la chaîne de confiance peut se briser. De plus, dans un écosystème scientifique de plus en plus décentralisé et avec l'adoption de nouvelles innovations (comme le partage de preprint et le modèle Publish, Review, Curate), les informations relatives à la version du document et à l'état d'avancement de son évaluation gagneront encore en importance, tout comme une plus grande transparence des évaluations par les pairs.

Intégration des ressources des dépôts dans les services de découverte

Bien que certains utilisateurs visitent directement les dépôts, la majorité du grand public consomment des informations via les moteurs de recherche, les assistants IA / chatbots, les agrégateurs d'actualités, les portails santé, etc. Dans ce scénario, les dépôts peuvent tout de même jouer un rôle important dans l'environnement épistémique au sens large en mettant en place des mesures permettant de retracer les affirmations circulant dans l'écosystème jusqu'à leur source d'origine dans une archive hébergée. Les dépôts peuvent ainsi servir de

plateformes de vérification soutenant les systèmes d'information utilisés au quotidien par les citoyens.

Cependant, les robots d'indexation génériques contournent souvent les protocoles de métadonnées en place dans les archives (COAR, 2025), ce qui signifie que les informations de provenance qu'elles contiennent ne se retrouvent pas toujours dans ces systèmes grand public. Pour prendre en charge ces scénarios, les dépôts devraient fournir des informations de provenance et d'autres informations contextuelles sous diverses formes facilement exploitables, à la fois par les machines et par les humains (en plus de maintenir des protocoles bien établis comme OAI-PMH). Par exemple, ces informations devraient être incluses sur les *landing pages*, où le HTML structuré peut être facilement exploré. Il conviendrait de mettre en œuvre des approches à la fois légères et efficaces, notamment le « signposting », afin de permettre une utilisation rapide et fluide des données structurées (Uthayakumar, 2026). Proposer diverses options adaptées aux outils automatisés pour extraire de l'archive les métadonnées pertinentes et d'autres informations sur la ressource augmente considérablement les chances que celles-ci apparaissent dans les interfaces de découverte.

En même temps, il existe aussi un sous-ensemble d'utilisateurs représentant des chercheurs d'information experts, tels que les praticiens, les enseignants, les étudiants ou les décideurs politiques, qui voudront toujours utiliser des outils axés sur le monde académique. Il est donc important de maintenir le soutien aux protocoles plus traditionnels comme OAI-PMH. Ces outils reposent généralement sur des agrégateurs de métadonnées intermédiaires tels que CORE, LA Referencia, OpenAIRE, et OpenAlex, qui réutilisent ensuite les données pour des services plus grand public. On peut citer pour exemples l'Observatoire des politiques en matière d'IA de l'OCDE - qui utilise les métadonnées d'OpenAlex pour produire des visualisations des tendances de la recherche en IA à destination des décideurs politiques et du grand public (OCDE.AI, 2026) - et le Portail européen d'information sur la santé (European Health Information Portal), qui collecte des métadonnées structurées à partir de sources de données santé dans toute l'Europe.

Comment les dépôts peuvent améliorer les requêtes des grands modèles de langage (LLM)

Les grands modèles de langage (LLM), qui commencent aujourd'hui à dominer la plupart des systèmes de découverte, introduisent un problème intrinsèque d'attribution qui rompt la chaîne de reconnaissance scientifique et de vérification (Peters & Chin-Yee, 2025). Lorsque vous posez une question à un LLM, il génère une réponse en synthétisant des schémas à partir des millions de documents sur lesquels il a été entraîné. Cependant, un LLM ne trace généralement pas quelles sources spécifiques ont contribué à quelles parties de sa réponse. Il n'existe donc aucun moyen fiable de relier les affirmations faites dans une réponse d'IA au matériel source original. Cependant, des développements récents suggèrent que des intégrations plus structurées entre les dépôts et les systèmes d'IA peuvent contribuer à résoudre ce problème lorsqu'elles sont appliquées strictement à la littérature scientifique. Le projet CORE-GPT a démontré que lorsque les systèmes d'IA sont configurés pour générer des réponses en utilisant des articles en accès ouvert plutôt que de s'appuyer uniquement sur des données d'entraînement, les citations fabriquées sont considérablement réduites (Pride et al., 2023). Asai et al. (2026) confirment cela à grande échelle avec OpenScholar, un modèle

enrichi par récupération utilisant 45 millions d'articles en accès ouvert, qui surpasse GPT-4o en exactitude. En revanche, GPT-4o a fabriqué entre 78 % et 90 % de ses références citées lorsqu'il fonctionnait sans accès à la littérature source. De même, d'autres outils populaires, tels que Gemini de Google ou Perplexity, commencent à prendre conscience que les utilisateurs souhaitent disposer de liens vers les sources originales et que cela renforce leur confiance, en particulier chez ceux qui recherchent des informations de manière plus approfondie. Dans ces cas, les dépôts qui exposent des informations de provenance fiables et des métadonnées structurées de manière adaptée aux machines permettront aux citations de devenir plus transparentes pour les systèmes d'IA, ce qui, à son tour, améliorera l'attribution des sources originales dans leurs services.

Conclusion

Les dépôts en libre accès sont depuis longtemps reconnus comme une infrastructure essentielle pour la communauté de la recherche, mais leur valeur s'étend bien au-delà du monde académique. Ils servent des publics très divers, allant des praticiens et enseignants aux entrepreneurs et animateurs associatifs des pays du Sud global, en offrant un accès sans barrière à des connaissances qui resteraient sinon verrouillées derrière des paywalls.

À un moment où la confiance du public dans l'information est sous une pression aiguë, les dépôts occupent une position particulièrement crédible. Leurs affiliations institutionnelles, leurs gouvernances transparentes et leur engagement à préserver les sources primaires en font des repères fiables dans un environnement informationnel de plus en plus turbulent. Pourtant, leur potentiel reste largement sous-exploité. Conçus principalement pour les communautés de recherche, ils n'ont pas encore pleinement adopté les fonctionnalités grand public qui leur permettraient d'atteindre leur impact sociétal maximal, comme les résumés en langage simple, la traçabilité structurée, et l'intégration transparente avec les outils de découverte que les utilisateurs utilisent au quotidien.

La croissance rapide des systèmes de découverte basés sur l'IA rend ce défi à la fois plus urgent et plus réalisable. Les dépôts qui investissent dans les métadonnées lisibles par machine, la provenance structurée, et l'interopérabilité avec les grands modèles de langage amélioreront non seulement l'attribution et réduiront le potentiel de désinformation, mais ils deviendront également fondamentaux pour un écosystème d'information plus digne de confiance.

En définitive, rendre les dépôts véritablement accessibles au grand public ne devrait pas être une réflexion après coup sur le plan technique. C'est une extension naturelle de la mission citoyenne que les universités et les institutions de recherche ont toujours eue. Les connaissances produites par la recherche financée par des fonds publics appartiennent au public. Les dépôts sont le chemin le plus direct pour honorer cet engagement.