Les dépôts et archives ouvertes libèrent le potentiel de la recherche multilingue

traduction en français de : Repositories unlock the potential of multilingual scholarship
https://zenodo.org/records/20718408

CCSD – June 2026

Le multilinguisme est une caractéristique essentielle d'un paysage de la communication scientifique sain, inclusif et véritablement mondial. Publier dans les langues locales permet aux populations de différents pays d'accéder aux travaux de recherche qu'elles financent et d'en tirer profit, tout en réduisant le désavantage structurel auquel sont confrontés les chercheurs qui ne travaillent pas en anglais. Les enjeux dépassent les seules carrières individuelles. Comme l'a indiqué COAR (2023), à la suite de l'Initiative d'Helsinki sur le multilinguisme dans la communication scientifique (2019) :

> « La disqualification des langues locales ou nationales dans l'édition scientifique est le facteur le plus important — et souvent oublié — qui empêche les sociétés de mobiliser et tirer parti des recherches menées sur leur territoire. »

Un système qui ne parle qu'une seule langue ne peut prétendre représenter les savoirs du monde entier.

Les dépôts contiennent des ressources dans de nombreuses langues

Les dépôts sont particulièrement bien placés pour favoriser la diffusion de la recherche multilingue, précisément parce qu'ils fonctionnent comme des nœuds gouvernés localement au sein d'un réseau mondial, plutôt que comme des plateformes centralisées optimisées pour un public unique. La plupart des interfaces des dépôts peuvent déjà être configurées pour fonctionner dans plusieurs langues et, dans de nombreux pays, une part importante des contenus déposés est publiée dans des langues autres que l'anglais. Il s'agit là d'un avantage structurel que les plateformes internationales et commerciales peuvent difficilement reproduire, car elles sont conçues pour maximiser l'engagement et la portée commerciale et sont, par conséquent, organisées autour de l'anglais comme langue par défaut. Les dépôts, en revanche, reflètent la réalité linguistique des communautés qu'ils desservent, ce qui en fait un socle plus équitable pour la communication scientifique mondiale et plus représentatif de sa diversité.

Bien que l'anglais ait longtemps été la langue de facto du discours scientifique international, cette prédominance est de plus en plus remise en question. Les gouvernements de

nombreuses régions réévaluent le rôle des langues locales dans la communication scientifique, sous l'effet de préoccupations liées à l'accès du public, à la pertinence culturelle et aux difficultés rencontrées par les chercheurs non anglophones. L'Amérique latine dispose depuis plusieurs décennies de son propre écosystème éditorial géré par la communauté, avec des revues régionales publiant principalement en espagnol et en portugais. Ce modèle suscite aujourd'hui un intérêt croissant. En Chine, par exemple, après avoir encouragé pendant des années les chercheurs à publier dans des revues internationales indexées dans le *Science Citation Index* (SCI), le gouvernement a commencé à inverser la tendance, en réduisant le recours à une évaluation exclusivement basée sur le SCI et en favorisant davantage la publication dans des revues chinoises et nationales (Owens, 2024 ; Yang et Huang, 2025). De même, le gouvernement fédéral canadien a investi dans le renforcement de l'écosystème de la recherche en langue française, en réponse aux données montrant que les chercheurs francophones sont confrontés à des désavantages structurels au sein de systèmes de publication et d'évaluation dominés par l'anglais (Patrimoine canadien, 2023 ; Beth et al., 2024). En Europe, la Coalition for Advancing Research Assessment (CoARA) a créé un groupe de travail sur le multilinguisme et les biais linguistiques, qui a élaboré un projet de cadre pour une évaluation de la recherche sensible aux enjeux linguistiques (Groupe de travail CoARA sur le multilinguisme et les biais linguistiques dans l'évaluation de la recherche, 2025). Ces évolutions témoignent d'une prise de conscience croissante : un système de communication scientifique construit presque exclusivement autour d'une seule langue n'est ni équitable ni, à long terme, optimal pour l'échange des connaissances.

Les dépôts figurent parmi les infrastructures les plus importantes pour le partage et la préservation des résultats de recherche multilingues. Plusieurs exemples majeurs illustrent l'ampleur de ce rôle. HAL, l'archive ouverte nationale française, est utilisée comme dépôt institutionnel par 98% des universités françaises et héberge plus de 1,5 million de documents en texte intégral, dont 27% sont en français (CCSD, 2026). Le Japan's National Institute of Informatics (NII) gère un service national d'hébergement et d'agrégation de dépôts contenant des milliers de documents en japonais. Au Brésil, l'IBICT coordonne un réseau national de dépôts consacré à la collecte et à la préservation des résultats de recherche en portugais. Au-delà de ces exemples nationaux, une étude de 2024 sur le paysage européen des dépôts, menée par COAR (Shearer et al., 2023), a révélé que les dépôts de tout le continent conservent systématiquement d'importantes collections dans leurs langues nationales. Cela suggère que la diversité linguistique est déjà une caractéristique des infrastructures des dépôts, et non une simple aspiration. Pris dans leur ensemble, ces exemples démontrent que les dépôts participent activement au maintien du caractère multilingue de la recherche scientifique mondiale.

Les systèmes de découverte actuels ne mettent pas en avant ce multilinguisme

Bien que les chercheurs du monde entier produisent des connaissances dans des centaines de langues, ces contenus restent largement invisibles pour les locuteurs non natifs. Lorsque nous cherchons ces connaissances, les outils dont nous disposons se comportent comme si seules quelques langues comptaient réellement. Même si les chercheurs et les autres

personnes à la recherche d'informations ne savent généralement lire qu'une ou deux langues, ils souhaitent connaître tous les travaux pertinents dans leur domaine, quelle que soit la langue de publication. Pourtant, les systèmes de découverte tels que Google Scholar présentent un biais systémique de leurs algorithmes à l'encontre des contenus non anglophones (Rovira et al., 2021), tandis que d'autres index scientifiques excluent tout simplement une grande partie des publications non anglophones de leurs services (Harzing, 2016). Les résultats de recherche non anglophones deviennent ainsi pratiquement invisibles. Parallèlement, la plupart des systèmes de découverte fonctionnent à l'intérieur de frontières linguistiques : les requêtes formulées dans une langue donnée renvoient principalement des résultats dans cette même langue. Cela crée des silos linguistiques, dans lesquels les recherches pertinentes produites dans d'autres langues restent effectivement invisibles — non pas parce qu'elles seraient dépourvues de pertinence, mais parce que les systèmes ne sont pas conçus pour relier les connaissances d'une langue à l'autre.

Le biais en faveur des contenus anglophones est également intégré aux systèmes d'évaluation de la recherche, de telle sorte qu'il incite fortement les chercheurs à publier en anglais. Comme les principaux services d'indexation et de citation ignorent largement les contenus non anglophones, les chercheurs qui souhaitent être reconnus dans des processus d'évaluation reposant sur ces index n'ont guère d'autre choix que de publier en anglais. L'Initiative d'Helsinki sur le multilinguisme dans la communication scientifique (2019) a averti que, si les systèmes de financement continuaient à ne récompenser que les publications en anglais, la communication scientifique multilingue ne se contenterait pas de stagner, mais serait structurellement érodée, créant ainsi un cercle vicieux dans lequel l'invisibilité conduirait à la perte de pertinence.

Ces déséquilibres structurels ont des conséquences épistémiques importantes. Konno et al. (2020) ont montré que l'omission d'études publiées dans des langues autres que l'anglais peut modifier de manière significative les conclusions des synthèses scientifiques, voire les inverser. En matière de citations, Amano et al. (2023) ont constaté que les références dans des langues autres que l'anglais représentent environ 65% des citations dans les rapports nationaux sur la biodiversité. Pourtant, elles ne représentaient que 3,4% des références dans les évaluations mondiales de l'Intergovernmental Science-Policy Platform on Biodiversity and Ecosystem Services. Lorsque la littérature non anglophone est systématiquement exclue, la base de données probantes utilisée pour les politiques publiques, les pratiques cliniques et l'action communautaire n'est pas seulement incomplète : elle peut également être trompeuse.

Ces déséquilibres structurels sont encore aggravés par l'intelligence artificielle. Les grands modèles de langage sont principalement entraînés sur des corpus en anglais et leurs performances diminuent considérablement dans les langues sous-représentées dans leurs données d'entraînement (Bengio et al., 2026). Comme ces modèles apprennent à partir d'environnements informationnels qui favorisent déjà très largement l'anglais, leurs résultats ne se contentent pas de refléter les hiérarchies linguistiques historiques : ils les renforcent et les prolongent activement. À mesure que l'IA s'intègre davantage dans la production, la découverte et la synthèse de la recherche, le risque est que des biais qui résultaient autrefois de décisions humaines soient encodés dans des systèmes apparemment neutres et, de ce fait, plus difficiles à remettre en question.

Placer le multilinguisme au cœur des biens communs de la connaissance mondiale

La justification historique de l'anglais comme langue par défaut de la science n'était pas dépourvue de logique. Un moyen de communication commun était réellement utile et l'anglais s'était imposé comme l'option la plus largement partagée entre les communautés de recherche. Mais cette justification relevait toujours d'un compromis pratique, et la technologie permet désormais de dépasser cette contrainte. Les grands modèles de langage offrent une qualité de traduction comparable à celle des systèmes dédiés, même pour des paires de langues sur lesquelles ils n'ont pas été explicitement entraînés (Hendy et al., 2023).

Les modèles d'intégration multilingues créent des espaces sémantiques communs à toutes les langues, permettant ainsi d'effectuer des recherches et d'extraire des informations en se basant sur le sens plutôt que sur la correspondance de mots-clés (Reimers & Gurevych, 2020 ; Matas, 2025). Les graphes de connaissances, quant à eux, permettent la résolution interculturelle des entités, c'est-à-dire l'identification d'un même concept, d'une même personne ou d'une même institution, quelle que soit la manière dont ils sont décrits dans différentes langues. Pris ensemble, ces progrès signifient que des contenus publiés dans n'importe quelle langue peuvent, en principe, être découverts et compris par une personne effectuant une recherche dans une autre langue. La barrière linguistique qui semblait autrefois rendre inévitable la domination de l'anglais est désormais un problème auquel il existe des solutions techniques.

Plusieurs initiatives visant à améliorer la découverte multilingue sont également menées dans des domaines spécifiques. En Europe, par exemple, l'infrastructure de recherche OPERAS gère GoTriple, une plateforme de découverte en sciences humaines et sociales qui facilite l'accès à des ressources multilingues provenant de différentes sources dans plusieurs pays. Par ailleurs, un modèle conceptuel de recherche sémantique multilingue, récemment développé au sein de la communauté COAR, propose que la recherche fondée sur le sens puisse fonctionner comme une couche complémentaire à la recherche par mots-clés (Matas, 2025). Plutôt que de rechercher des mots identiques, cette approche saisit le sens sous-jacent des documents et des requêtes, ce qui permet d'effectuer des recherches basées sur la similarité sémantique entre différentes langues.

Ainsi, une requête formulée dans une langue peut faire apparaître des travaux pertinents produits dans une autre, sans que les utilisateurs aient besoin d'anticiper ou de traduire les mots-clés. Le modèle propose de rendre cette fonctionnalité accessible par les mêmes mécanismes d'interopérabilité que ceux évoqués précédemment, afin que les institutions puissent adopter la technologie en fonction de leurs ressources. Bien que cette approche demeure exploratoire, les premières preuves de concept ont produit des résultats encourageants. Elles montrent notamment que les langues moins dominantes peuvent gagner en visibilité.

Un système mondial de découverte multilingue nécessitera un réseau cohérent et distribué de fournisseurs de données à travers le monde, utilisant des normes et des protocoles communs pouvant être indexés à grande échelle tout en préservant la gouvernance locale et

la diversité linguistique. Ces systèmes devront pouvoir relier les requêtes et les contenus par-delà les langues, plutôt que de fonctionner dans un seul contexte linguistique.

Les dépôts sont idéalement placés pour constituer une partie intégrante de ce système, non seulement en raison des contenus qu'ils conservent, mais aussi de leur conception même : ce sont des systèmes interopérables et distribués, capables d'exposer des métadonnées structurées à grande échelle.

Contrairement aux grands index scientifiques, dont la couverture repose principalement sur des publications issues de revues internationales et demeure fortement biaisée en faveur des contenus anglophones, les dépôts existent dans toutes les régions du monde et reflètent une véritable diversité géographique et linguistique. Un système de découverte construit à partir des dépôts offrirait donc une vision plus complète de l'environnement mondial de la recherche multilingue que celle qui est actuellement disponible.

Les prérequis techniques d'un tel système existent déjà au sein du réseau mondial des dépôts. L'interopérabilité constitue un principe essentiel de leur conception : les dépôts adoptent des normes et des protocoles communs — notamment OAI-PMH — qui permettent d'exposer les contenus de manière cohérente et lisible par les machines, entre différentes plateformes et organisations.

Toutefois, une véritable découverte multilingue nécessitera l'adoption de pratiques supplémentaires par les dépôts. En 2023, le groupe de travail de COAR *Supporting Multilingualism and non-English Content in Repositories* a publié 13 recommandations à l'intention des dépôts afin d'améliorer la visibilité de leurs contenus. Des progrès ont été réalisés dans l'adoption de ces recommandations, mais de nombreux dépôts ne les ont pas encore pleinement mises en œuvre et elles ne sont pas intégrées dans la plupart des processus de curation. Il reste donc nécessaire de poursuivre les efforts pour commencer à résoudre ce problème.

Concrétiser cette vision à la vitesse et à l'échelle requises exigera une coordination et un engagement accrus de la part de la communauté des dépôts. De par leur conception, les dépôts sont distribués et largement autonomes, leur coordination reposant sur la recherche d'un large consensus plutôt que sur une gouvernance centralisée. Cela implique non seulement l'adoption de normes communes, mais aussi un meilleur alignement des pratiques de gestion des métadonnées, de description multilingue et d'exposition des contenus en vue de leur traitement par les machines. Ces difficultés ne sont toutefois pas insurmontables. Elles relèvent du même type de défis d'action collective que ceux que la communauté a déjà surmontés, notamment grâce à l'adoption généralisée de l'OAI-PMH et des recommandations de COAR pour les dépôts de nouvelle génération. Dans les prochains mois, COAR mènera plusieurs activités afin de renforcer la coordination de la communauté, notamment au moyen d'un éventuel projet pilote. Elle intensifiera également ses efforts pour promouvoir les bonnes pratiques qui permettront d'accroître la valeur des contenus multilingues conservés dans le réseau.

Conclusion

La prédominance de l'anglais dans la communication scientifique n'a jamais été un choix délibéré. Il s'agissait d'une solution pratique qui, au fil du temps, s'est muée en une habitude si profondément ancrée qu'elle en est devenue presque imperceptible. Les conséquences de cette habitude sont aujourd'hui bien documentées : des recherches produites dans des centaines de langues restent méconnues, les données probantes servant de base aux politiques et aux pratiques sont systématiquement incomplètes, et les chercheurs et les communautés dont les connaissances sont exclues subissent des pertes rarement prises en compte par ceux qui profitent du système actuel. À mesure que l'IA s'intègre plus profondément dans la manière dont la recherche est trouvée et utilisée, le risque est que ces déséquilibres deviennent computationnels, encodés dans des systèmes qui semblent objectifs et sont donc beaucoup plus difficiles à contester.

Ce qui a fait défaut jusqu'à présent, ce n'est ni la technologie, ni les contenus, ni même les normes techniques. C'est la reconnaissance du fait que le multilinguisme n'est pas une préoccupation secondaire, mais une condition nécessaire des biens communs mondiaux de la connaissance. La bonne nouvelle est que l'infrastructure nécessaire pour construire un meilleur système existe déjà. Les dépôts, répartis dans toutes les régions du monde et gouvernés par les communautés qu'ils servent, abritent de vastes collections de travaux de recherche multilingues. Ils reposent sur des normes d'interopérabilité qui rendent techniquement possibles l'indexation et l'exploration à grande échelle. Ils commencent également à adopter les pratiques nécessaires à un système de découverte véritablement inclusif.

Les dépôts offrent une voie crédible vers un système véritablement différent de communication scientifique : un système qui reflète une diversité linguistique beaucoup plus grande, qui est gouverné par les communautés productrices de connaissances et qui est suffisamment résilient pour rester ouvert, quelles que soient les pressions commerciales ou géopolitiques. La réalisation de ce potentiel nécessitera des normes coordonnées et une volonté collective de considérer le multilinguisme non comme un ajout, mais comme un principe fondateur. Les arguments en faveur de l'action sont solides. Les outils sont disponibles. Il faut désormais les mettre en œuvre.

Remerciements

Ce document a été rédigé par Diana Furcila, consultante COAR, et Kathleen Shearer, directrice générale de COAR, puis relu par Lautaro Matas et le comité exécutif de COAR.