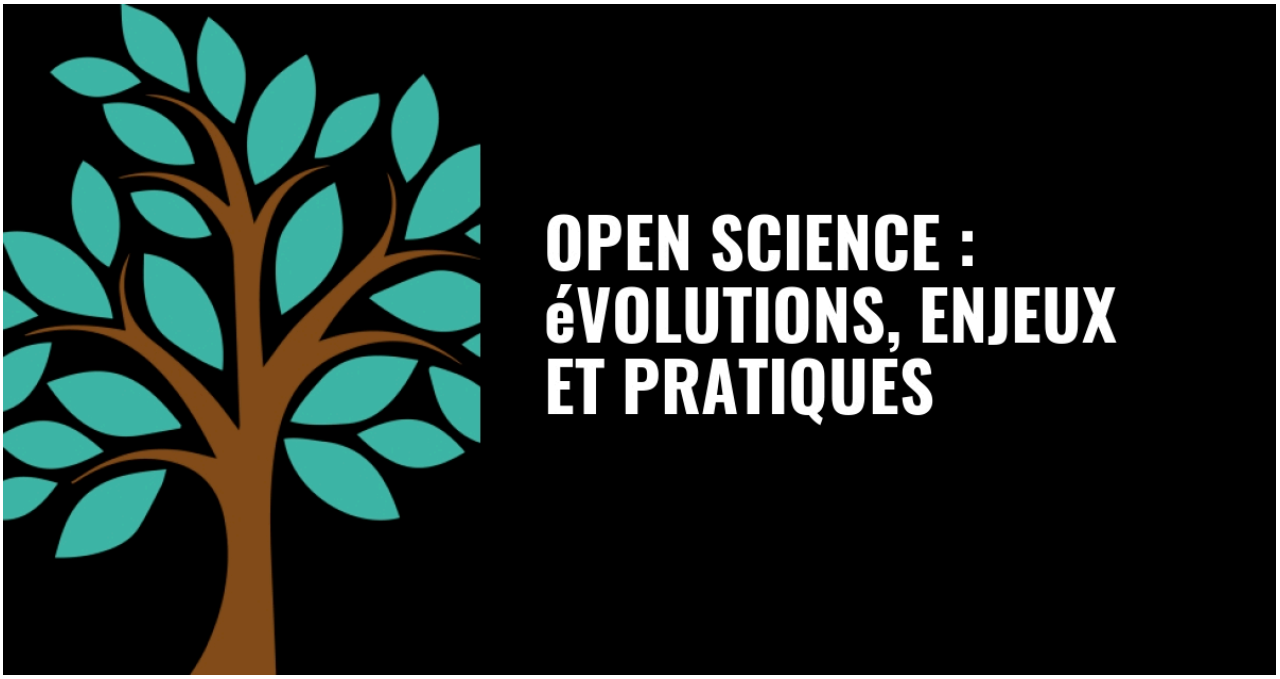
Construire un véritable écosystème d'Open Data en santé : défis et perspectives en Amérique latine

openscience.pasteur.fr/2025/12/08/construire-un-veritable-ecosysteme-dopen-data-en-sante-defis-et-perspectives-en-amerique-latine/

CeRIS - Institut Pasteur 8 décembre 2025



[Version espagnole accessible ici : « Construir un verdadero ecosistema de Open Data en salud: desafíos y perspectivas en América Latina »]

Les sciences biomédicales génèrent aujourd'hui d'immenses quantités de données complexes et coûteuses. Pourtant, une grande partie de ces informations reste inexploitée. Les États-Unis et l'Europe ont mis en place des politiques d'open data afin de garantir que les ressources publiques se transforment en connaissances accessibles et utiles, conformément aux principes de l'UNESCO. Mais entre la théorie et la pratique, un fossé persistant demeure : la fameuse formule des données disponibles « sur demande raisonnable » est souvent tournée en dérision, car accéder à un jeu de données – même totalement anonymisé – reste un parcours frustrant. En 2022, une étude a montré que 93% des auteurs contactés n'avaient pas partagé les données annoncées comme « disponibles ».

Et pourtant, lorsque les données sont réellement accessibles, les résultats peuvent être transformateurs. Au début de la pandémie de COVID-19 au Brésil, des chercheur·e·s ont identifié des éléments clés chez des patient·e·s présentant des comorbidités en utilisant uniquement des données publiques déjà disponibles, démontrant ainsi que l'ouverture peut accélérer les découvertes et produire des connaissances utiles.

En Amérique latine, les compétences scientifiques existent bel et bien, mais le manque d'accès et d'infrastructures freine largement leur potentiel. Une étude récente menée dans quatre universités à Cuba, au Pérou et en Bolivie a révélé que 61 % des données

quantitatives produites dans le cadre de projets de recherche finissent stockées sur des ordinateurs personnels ; à peine 30 % sont conservées sur des serveurs institutionnels. Pire encore, plus de la moitié des personnes interrogées ont déclaré avoir perdu des données au moins une fois. Beaucoup reconnaissent une faible connaissance des règles éthiques ou légales, notamment lorsqu'il s'agit de données sensibles, et soulignent l'insuffisance, voire l'absence, d'infrastructures institutionnelles pour préserver, partager ou réutiliser les données.

À titre personnel, cette étude m'a rappelé mon expérience au Mexique, pendant mon master (autour de 2016) : je ne me souviens d'aucun espace institutionnel dans « le cloud » où nous aurions pu stocker les données de nos projets. Les données restaient presque toujours sur nos ordinateurs personnels ou, au mieux, dans des services externes de type Dropbox, gérés individuellement. Presque dix ans plus tard, j'ignore si la situation a évolué, mais j'espère sincèrement que l'infrastructure s'est améliorée et que les pratiques de gestion des données se sont professionnalisées.

Cette réflexion nous rappelle que rendre les données accessibles ne dépend pas seulement de politiques globales ou de bonnes intentions : cela exige que les chercheur·e·s disposent d'infrastructures fiables, d'un soutien institutionnel, de connaissances juridiques et d'une culture partagée de gestion responsable. Sans correction de ces lacunes structurelles, beaucoup de données resteront perdues, vulnérables ou inaccessibles, et avec elles, des découvertes potentielles, des collaborations et des avancées scientifiques resteront en chemin.

Le tableau que je viens de décrire concernant la gestion des données dans de nombreuses universités latino-américaines soulève une autre question : si les chercheur·e·s peinent déjà à conserver et sécuriser leurs propres données, quels moyens ont-ils pour les partager et les mettre à disposition de la communauté scientifique ? En d'autres termes : les jeux de données produits par la recherche en Amérique latine sont-ils véritablement ouverts et réutilisables ?

Une méta-analyse récente a été menée pour répondre à ce manque structurel : comme la majorité des bases de données publiques en santé proviennent de pays à revenu élevé, on sait encore très peu de choses sur ce qui est réellement disponible en Amérique latine. L'étude s'est donné pour objectif de cartographier de manière systématique les jeux de données de santé ouverts dans la région : combien existent, de quels pays ils proviennent, quelles modalités et formats ils utilisent, et dans quelle mesure ils sont effectivement accessibles et réutilisables.

Les auteurs ont analysé les publications de 2006 à 2023 afin d'identifier les jeux de données en santé sur la population latino-américaine disponibles en accès libre. Sur plus de 700 publications initiales, 141 études ont été retenues, et de celles-ci ont été extraits des renseignements sur l'origine, le type, le format et l'accessibilité des données.

Les résultats sont révélateurs : 61 jeux de données publics provenant de 23 pays ont été identifiés. Même si ce chiffre peut sembler encourageant, la distribution et la nature de ces données montrent de fortes limitations. La majorité des travaux reposent sur des bases de données de santé bien établies, en particulier DATASUS au Brésil. En revanche, très peu d'études ont généré et partagé leurs propres données : sur les 141 articles analysés, seuls 23 ont créé de nouveaux jeux de données ouverts.

Quant à la nature des données, la domination est nette : 88,7 % des études utilisent des données tabulaires (informations épidémiologiques, démographiques ou de santé publique). Les données plus complexes, telles que les images médicales, la génomique, les signaux, les textes cliniques ou les dossiers médicaux, sont extrêmement rares ; seules quelques études incluent des images, et encore moins combinent plusieurs modalités.

Ce panorama a des conséquences majeures. Le fait que la majorité des données ouvertes de santé en Amérique latine proviennent d'un seul pays (le Brésil) et qu'elles soient essentiellement agrégées limite fortement la représentativité régionale. Les modèles ou études basés sur ces données refléteront difficilement la diversité épidémiologique, sociale, génétique ou clinique de toute la région. Les auteurs soulignent qu'avec les données actuellement disponibles, de nombreuses questions essentielles — variabilité clinique selon le pays, spécificités locales, inégalités en santé — ne peuvent tout simplement pas être abordées.

En outre, l'étude rappelle qu'il ne suffit pas d'avoir des données : encore faut-il qu'elles soient bien documentées, accessibles et prêtes à être réutilisées (selon les principes FAIR, qui garantissent des données trouvables, accessibles, interopérables et réutilisables). Sans standards, sans métadonnées solides et sans dépôts fiables, même les données « ouvertes » risquent de rester invisibles ou inutilisables.

Enfin — et c'est, à mon sens, l'un des points les plus cruciaux — les auteurs lancent un appel à l'action. Ils recommandent d'investir dans les infrastructures, de développer des politiques institutionnelles d'open data, de créer des dépôts pérennes, de mettre en place des protocoles d'anonymisation et de gouvernance éthique, et de créer des incitations au partage de données. Car une véritable science ouverte exige bien plus que des "demandes raisonnables" : elle nécessite des systèmes clairs, des métadonnées complètes et un engagement réel en faveur de la mise à disposition des données. Chaque jeu de données rendu accessible est une opportunité pour formuler de nouvelles questions, renforcer la recherche et consolider un écosystème de recherche plus solide, plus équitable et réellement représentatif de l'Amérique latine.

María Gutiérrez Sánchez, postdoctorante à l'Institut Pasteur