Choisir le meilleur entrepôt de données

1. Qu'est-ce qu'un entrepôt de données de recherche

Un entrepôt de données de recherche (Research Data Repository ou Data Repository) est une infrastructure destinée à accueillir, conserver, rendre visibles et accessibles des données de recherche. Un entrepôt se distingue d'un catalogue, par sa capacité à assurer l'hébergement, la gestion et la curation des données et pas uniquement le système d'information (catalogage et exposition des métadonnées moissonnées à partir d'autres structures).

Son rôle est de permettre le dépôt de données, leur description, leur accès et leur partage en vue de leur réutilisation. Chaque entrepôt dispose généralement d'une politique de dépôt, de description et de diffusion des données.

Ces entrepôts s'inscrivent dans une démarche de partage et d'ouverture des données selon les principes FAIR pour que les données soient "Faciles à trouver, Accessibles, Interopérables et Réutilisables" (en anglais : *Findable, Accessible, Interoperable, Reusable*).

2. Pourquoi déposer des données dans un entrepôt

Rendre accessibles ses données dans un entrepôt est une des pratiques majeures de la Science Ouverte. Cela contribue à la transparence des processus de recherche et à la reproductibilité des résultats publiés dans des articles scientifiques. Partager ses données permet aussi à d'autres équipes de les réutiliser sans avoir à les générer une nouvelle fois, ce qui permet un gain de temps et assure une meilleure utilisation des fonds de la recherche.

Déposer ses données dans un entrepôt assure leur préservation, leur visibilité et leur accès et facilite aussi leur partage, réutilisation et citation grâce à l'attribution par l'entrepôt d'un identifiant numérique unique et pérenne à chaque jeu de données. En outre, le crédit reçu en tant qu'autrice, auteur, contributrice ou contributeur des données diffusées, permet d'accroître sa notoriété.

Mettre à disposition ses données répond aux attentes des financeurs. En effet, si vous publiez des articles de recherche au cours de votre projet, les financeurs recommandent que les données qui sont à l'origine des résultats soient déposées dans un entrepôt.

Les éditeurs scientifiques ont aussi de plus en plus d'exigences sur le partage des données (et codes sources) à l'origine des articles de recherche. Dans ce but, les revues recommandent généralement une sélection d'entrepôts. Dans les cas où existe un entrepôt disciplinaire reconnu par la communauté scientifique, il pourra être obligatoire de l'utiliser (ex : génomique, cristallographie, mycologie). Le lien vers l'entrepôt et l'identifiant numérique attribué au jeu de données seront renseignés dans le paragraphe *Data Availability Statement* de l'article de recherche.

Déposer ses données dans un entrepôt apporte ainsi de nombreux avantages :

- conservation des données dans un environnement sécurisé
- visibilité des données et accès facilité pour les moteurs de recherche
- interopérabilité des données grâce à l'utilisation de standards de métadonnées
- découverte, réutilisation et citation du jeu de données facilitées par son identifiant pérenne
- gestion des modalités de partage des données par l'attribution de licences de diffusion
- publication facilitée, conforme aux principes de science ouverte
- respect des recommandations des financeurs et institutions sur l'ouverture des données
- reproductibilité de la recherche, intégrité et validation scientifique améliorées
- valorisation des données par leur réutilisation dans de nouvelles études et innovations.

3. Différentes catégories d'entrepôts de données

Il existe plusieurs catégories d'entrepôts de données : internationaux, nationaux ou institutionnels ; génériques, disciplinaires ou thématiques.

Certaines disciplines sont, en effet, bien structurées avec de nombreux entrepôts disciplinaires reconnus (ex : biologie, sciences de l'environnement) ; d'autres sont moins bien dotées (ex : agronomie). Lorsque des entrepôts thématiques ou disciplinaires existent, les communautés scientifiques, les éditeurs, les financeurs recommandent de les utiliser.

Pour accroitre la visibilité et la réutilisation des données, les entrepôts thématiques et disciplinaires complètent les métadonnées (informations sur les données) basiques pour tout document (ex : titre, sujet, créateurs, contributeurs, date, …) par des métadonnées disciplinaires spécifiques. Ces métadonnées, souvent sous forme standardisée, ajoutent des éléments descriptifs spécifiquement adaptés à chaque discipline (ex : couvertures géographique, temporelle et taxonomique, spécifications d'un séquenceur, conditions de réalisation d'une enquête). L'utilisation de standards de métadonnées et de bonnes pratiques disciplinaires pour documenter et mettre en forme les données, facilitent la découverte, la compréhension, l'interprétation des données et donc leur réutilisation.

Exemples d'entrepôts de données :

- Thématiques ou disciplinaires : Data-Terra (système terre), ENA ou GenBank (séquences génétiques), UniProt (protéines), TRY (caractères botaniques), GBIF (biodiversité), Pangaea (sciences de la terre et de l'environnement), WormBase (nématologie), Movebank (mobilité animale), West African Vegetation, DataFirst (enquêtes socio-économiques en Afrique), Protocols.io (protocoles), etc.
- Institutionnels (en France) : Cirad Dataverse,, DataSuds (IRD)
- National : Recherche Data Gouv: la plateforme française fédérée des données de recherche, créée dans le cadre du Plan national pour la science ouverte
- Génériques : Zenodo, Dryad, Figshare
- Liés à un éditeur ou une revue (tout en restant un entrepôt ouvert) : Mendeley (Elsevier), Harvard Dataverse (Ubiquity Press), Harvard Dataverse (*Economics)* ou GigaDB (GigaScience).

4. Les questions à se poser avant de déposer des données dans un entrepôt

Le dépôt de données dans un entrepôt a pour objectif de partager des données, c'est-à-dire de les rendre accessibles pour qu'elles puissent être réutilisées par d'autres, que ce soit des scientifiques, des entreprises, des décideurs, des ONGs ou des citoyennes et citoyens.

La décision de rendre publiques des données de recherche (Voir la fiche CoopIST : Ouvrir ses jeux de données scientifiques) s'appuie sur des critères scientifiques, réglementaires, juridiques, humains, économiques et techniques et implique l'ensemble des contributeurs et partenaires d'un projet.

Les questions à se poser sont notamment :

- Quelles sont les obligations d'ouverture des données qui s'appliquent ? L'obligation peut être imposée par le financeur du projet, par une loi nationale, européenne ou internationale, par la politique des données de certains partenaires, par la revue dans laquelle vous publiez, etc. Le principe est « aussi ouvert que possible, aussi fermé que nécessaire » ; si des raisons justifient de ne pas ouvrir les données, elles ne seront pas rendues publiques.
- Quelle est la valeur scientifique des données et leur potentiel de réutilisation ? L'intérêt et l'utilité actuelle ou future, scientifique, environnementale, économique, patrimoniale ou sociale, des données peuvent guider le choix.
- Avez-vous le droit de rendre publiques ces données ? En d'autres termes, avez-vous respecté :
  les droits de propriété intellectuelle ?
  Ex : données obtenues en partenariat ou contenant des images protégées par le droit d'auteur
  les obligations contractuelles ?
  Ex : utilisation de données préexistantes, issues d'un projet précédent ou téléchargées à partir d'un entrepôt (ex *: FAOSTAT, GBIF, CHELSA, WorldPop*), éventuellement protégées par des droits spécifiques ou des licences
  les réglementations juridiques ou éthiques ?
  Ex : données personnelles collectées lors d'enquêtes et qui doivent être supprimées (anonymisation) pour respecter les droits des personnes
  Ou données issues de ressources génétiques ou de savoirs traditionnels associés qui nécessitent de respecter la Réglementation APA sur l'accès et le partage des avantages
  Ou données qui soulèvent des questions éthiques (ex : expérimentation animale, essais cliniques chez l'homme, recherches ayant un impact sur l'environnement, etc.) et requièrent la validation par un comité d'éthique
- Avez-vous obtenu l'accord de tous les contributeurs ?
- Avez-vous évalué le temps et l'effort nécessaires à la mise en forme des données et des métadonnées pour répondre aux exigences de l'entrepôt ? Si vous avez contribué et suivi le Plan de gestion des données de votre projet (Voir la fiche CoopIST : https://doi.org/10.18167/coopist/0066) alors vos données sont quasi prêtes et leur dépôt en sera facilité.
- Avez-vous défini les conditions de réutilisation des données que vous avez produites ? Souhaitez-vous exclure une utilisation commerciale de vos données ? ou que vos données soient accessibles à tous avec juste l'obligation de vous citer si elles sont réutilisées ? ce dernier cas est le plus recommandé et correspond à la licence CC-BY. Le plus souvent, l'entrepôt de données propose un choix de licences de diffusion (Voir la fiche CoopIST : https://doi.org/10.18167/xtnv-d457).

5. Comment choisir un entrepôt de données

Le meilleur guide pour le choix d'un entrepôt de données sera souvent de suivre les pratiques de votre communauté scientifique. En effet, il est recommandé de choisir un entrepôt thématique ou disciplinaire plutôt que généraliste ; vos données seront mieux mises en valeur. En revanche, si aucun

entrepôt thématique n'est recommandé par votre communauté, il est conseillé de déposer vos données dans l'entrepôt de votre établissement s'il existe, ou sinon dans un entrepôt pluridisciplinaire, national (ex : Recherche Data Gouv en France) ou international (ex : Zenodo pour l'Europe et même au-delà ou DataFirst pour l'Afrique). La politique du financeur ou de votre établissement peut également orienter votre choix ; de même les instructions aux auteurs de la revue dans laquelle vous soumettrez votre article peuvent aussi indiquer des entrepôts spécifiques.

En 2023, un guide méthodologique pour sélectionner un entrepôt thématique de confiance a été publié par le Collège des données de la recherche, à la demande du Ministère de l'Enseignement supérieur et de la Recherche : https://doi.org/10.52949/52.

Il existe des listes et répertoires d'entrepôts de données pour vous aider à choisir :

- re3Data : répertoire global et gratuit d'entrepôts de données créé en 2012 par le consortium international DataCite qui œuvre pour le partage des données de recherche. re3data renseigne sur 3240 entrepôts (aout 2024), les institutions qui les portent, les licences proposées, les standards utilisés, les conditions de dépôt de jeux de données, ... re3data est recommandé par les éditeurs et les financeurs.
- Disciplinary repositories : liste couvrant de nombreux domaines dont Agriculture, Économie, … crée par Simmons University (USA).
- Fairsharing : guide sur les normes et standards, les entrepôts et bases de données, les politiques et recommandations en termes de données et métadonnées.
- ELIXIR Deposition Databases for Biomolecular Data : liste d'entrepôts en science de la vie proposée par l'organisation européenne ELIXIR.
- Cat OPIDoR : wiki des services dédiés aux données de recherche, hébergé par le CNRS. Cat OPIDoR référencie 213 entrepôts de données en France en aout 2024.

Le choix d'un entrepôt peut être facilité en regardant dans quels entrepôts sont déposées des données similaires. Pour cela vous pouvez utiliser des moteurs de recherche tels que DataciteCommons ou Google Dataset Search (voir la fiche CoopIST : Trouver des jeux de données via des bases pluridisciplinaires et des moteurs de recherche ou la page Explorer les moteurs de recherche scientifiques).

Pour sélectionner votre entrepôt, vérifiez qu'il répond aux critères suivants :

- adapté au type de données que vous allez déposer
- adapté à la taille des fichiers que vous allez déposer (certains entrepôts ont des limites de taille de fichier individuel et de taille totale du dépôt)
- répondant aux recommandations du bailleur de votre projet, de votre institution, ou de la revue dans laquelle vous publiez
- reconnu dans votre discipline et par la communauté scientifique (certains entrepôts, encore peu nombreux, sont certifiés CoreTrustSeal)
- attribuant automatiquement un identifiant numérique pérenne, univoque à chaque jeu de données (Voir la fiche CoopIST : Identifier et rechercher une publication ou un jeu de données par son DOI)
- assurant la conservation des données, c'est-à-dire leur pérennité
- gratuit (la plupart des entrepôts) ou pratiquant des coûts de dépôt de données acceptables
- proposant les modalités d'accès aux données, adaptés à vos besoins : accès libre, après enregistrement, accès restreint, sur demande, différé par un embargo

- attribuant la licence de diffusion des données adaptée à vos exigences (Voir la fiche CoopIST : Connaître et utiliser les licences Creative Commons). Attention : 1) certains entrepôts imposent une licence alors que d'autres proposent un choix de licences et 2) lorsque les données sont liées à un article scientifique, la licence de diffusion appliquée aux données doit aussi répondre aux exigences de la revue (consultez les instructions aux auteurs).

Liens utiles

Science ouverte - Données de la recherche. Guide de la collection Passeport pour la science ouverte. 2024. https://www.ouvrirlascience.fr/science-ouverte-donnees-de-la-recherche/

Partager les données liées aux publications scientifiques – Guide pour les chercheurs. 2022. https://www.ouvrirlascience.fr/partager-les-donnees-liees-aux-publications-scientifiques-guide-pour-les-chercheurs/

Science ouverte - Codes et Logiciels. Guide de la collection Passeport pour la science ouverte. 2022. https://www.ouvrirlascience.fr/science-ouverte-codes-et-logiciels/

Passeport pour la Science Ouverte. Guide pratique à l'usage des doctorantes et des doctorants. Guide pratique à l'usage des doctorantes et des doctorants. Collection Passeport pour la Science Ouverte. 2ème édition 2024. https://www.ouvrirlascience.fr/passeport-pour-la-science-ouverte-guide-pratique-a-lusage-des-doctorants/

Sélectionner un entrepôt thématique de confiance pour le dépôt de données : méthodologie et analyse de l'offre existante. Comité pour la Science Ouverte. 2024. https://hal.science/ESPACE-DEV/hal-04534321v1.

Biard Yannick ; Dedieu Laurence

Délégation à l'information scientifique et à la science ouverte

Janvier 2025

Mise à jour de la fiche : Dedieu, L. ; Barale, M. 2020. Déposer des données dans un entrepôt, en 6 points. Montpellier (FRA) : CIRAD, 4 p.

Comment citer ce document :

Biard, Y. Dedieu, L. Choisir le meilleur entrepôt de données. 2025. Montpellier (FRA) : CIRAD, 5 p. *https://doi.org/10.18167/coopist/0070*