# Les entrepôts de données

openscience.pasteur.fr/2025/05/23/les-entrepots-de-donnees/

CeRIS - Institut Pasteur 23 mai 2025

*[Mise à jour d'un article initialement publié en janvier 2021]*

Les entrepôts de données de recherche sont des services en ligne permettant le stockage de données scientifiques et de leurs métadonnées descriptives. Leur objectif premier est de faciliter la **diffusion** et la **découverte** des données de la recherche.

Un « jeu de données » déposé sur un entrepôt se présente sous la forme d'une **page web** (un exemple ici) sur laquelle on retrouve :

- un **identifiant,** soit un identifiant unique et pérenne (DOI par exemple), soit un identifiant unique (souvent appelé *Accession Number*),
- des **métadonnées** décrivant le jeu de données (titre, auteurs, description, type de données, méthode, publication associée…),
- des informations sur les **conditions d'accès et de réutilisation** des données (accès ouvert, sur demande, licence de diffusion…),
- (facultatif) un ou plusieurs **fichiers,** correspondant aux données elles-mêmes et/ou à des fichiers de documentation (fichier README par exemple).

Cette page, facile à trouver sur le web, permet au scientifique de **faire connaître l'existence du jeu de données à un large public**. Mais les fichiers de données eux-mêmes ne sont pas systématiquement accessibles librement :

- La majorité des entrepôts sont des **entrepôts ouverts** : sur la page web de chaque jeu de données, les fichiers peuvent être téléchargés librement et réutilisés selon les conditions définies par la licence de diffusion (un exemple ici),
- Certains entrepôts permettent aux scientifiques de **contrôler l'accès aux fichiers** : la page web expose les métadonnées décrivant le jeu de données et indique la procédure à suivre pour demander l'accès aux fichiers (un exemple ici).

Notons que la plupart des entrepôts **ne sont pas conçus pour assurer la pérennité sur le long terme des données**. Il arrive également que certains entrepôts disparaissent sans nécessairement adopter de stratégie pour éviter la perte de données. Ils ne devraient donc pas être utilisés pour l'archivage des données sur le long terme.

Il existe de très nombreux entrepôts de données (plus de 1900 en sciences de la vie d'après Re3data) et ils peuvent être catégorisés selon deux grands types :

- les entrepôts **disciplinaires** ou **thématiques,** en imagerie, chimie, neuroscience, protéomique… (retrouvez ici trois exemples d'entrepôts en neuroscience, et ici une liste d'entrepôts thématiques de confiance publiée par le Comité pour la science ouverte) ;

- les entrepôts **généralistes** ou **pluridisciplinaires**, ouverts à tous types de données (retrouvez ici un comparatif de quatre entrepôts généralistes).

Chaque entrepôt a des caractéristiques et des fonctionnalités différentes qu'il est important de prendre en compte avant de faire son choix. Retrouvez ici nos conseils pour trouver un entrepôt adapté à vos besoins.