



UNIVERSITÉ DE
MONTPELLIER

Trouver, réutiliser et déposer des données de la recherche : les entrepôts de données

Webinaire du 21 juin 2022

Equipe d'Appui à la Science Ouverte de l'UM



— Objectifs et plan schématique de la formation —

• Objectifs

Savoir trouver, réutiliser et déposer des données de la recherche dans le contexte de la Science Ouverte

Axes du webinaire :

1. **Trouver des données de la recherche : data papers, moteurs de recherche spécialisés, entrepôts et annuaires d'entrepôts de données**
2. **Réutiliser des données : les différentes licences**
3. **Déposer des données en accord avec les principes FAIR**
4. **Qu'en est-il de l'archivage ?**

— 1. Trouver des données de la recherche

Pour trouver des données de la recherche, il existe **trois possibilités** :

- La **bibliographie des articles** (les DOI mentionnés renvoient vers les données utilisées)
- Les **data papers** (les revues spécialisées proposent de faire de la veille via des flux RSS ou des alertes)
- Les **entrepôts, répertoires** d'entrepôts ou **moteurs de recherche** de données

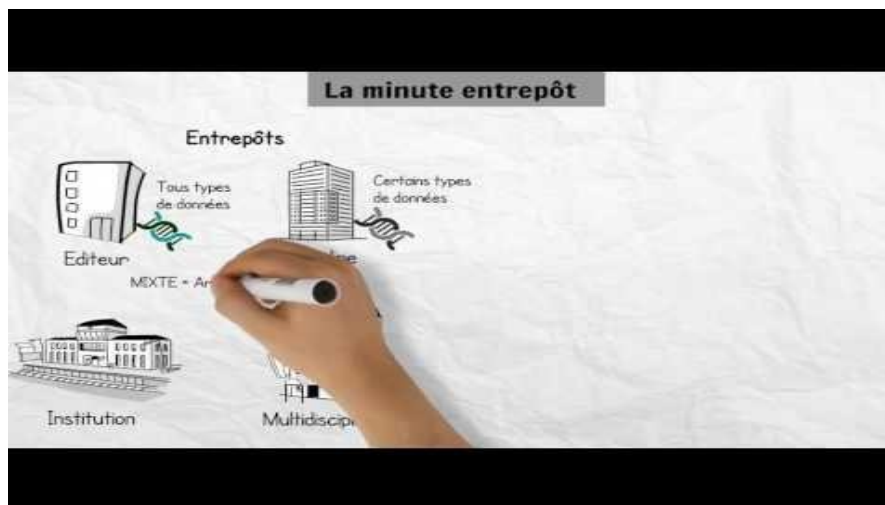
— 1. Trouver des données de la recherche : les entrepôts

Un entrepôt est un service en ligne permettant le dépôt, la description, la conservation, la recherche et la diffusion des jeux de données et des codes sources.

Ils peuvent être **généralistes ou thématiques**, publics ou propriétaires.

Fonctionnalités :

- Moteurs de recherche et **filtres** pour faciliter l'accès aux données.
- Les interfaces de recherches peuvent être plus ou moins avancées selon les entrepôts.



source : <https://doranum.fr/depot-entrepots/minute/>

— 1. Trouver des données de la recherche : la typologie des entrepôts

Disciplinaires

- **NAKALA: entrepôt national français (SHS)**
- SEANOE: entrepôt national français (sciences marines)
- PANGAEA (sciences de la Terre et de l'environnement)
- Réseau Quetelet: entrepôt national français (sciences sociales)

Multidisciplinaires

- **ZENODO: entrepôt mixte (publications scientifiques et jeux de données) – tous types de données**
- Figshare: tous types de données
- Dryad: données sous-jacentes aux publications
- Recherchedatagouv : tous types de données et codes sources, entrepôt national français
- Github/Gitlab : pour les codes sources

Institutionnels

- Edinburgh DataShare
- Merritt
- Enlighten : Research Data
- Dataverse CIRAD

— 1. Trouver des données de la recherche : les annuaires

Les annuaires d'entrepôts :

- Annuaires multidisciplinaires :
 - [Re3data](#)
 - [OAD](#) (Open Access Directory)
- Annuaires spécialisés :
 - [FAIRsharing](#) (sciences de la vie et biomédecine)
 - [NIH Data Sharing Repositories](#) (biomédecine)
 - [CAT Opidor](#) (entrepôts de données français)
- Annuaires mixtes :
 - [OpenDOAR](#)
 - [ROAR](#)
- Annuaire d'entrepôts certifiés :
 - [CoreTrustSeal](#)



OpenDOAR



— 1. Trouver des données de la recherche : les entrepôts certifiés

CoreTrustSeal est un annuaire d'entrepôt donnant accès à une liste **d'entrepôts certifiés** :

- Organisme communautaire sans but lucratif qui promeut le développement d'infrastructures de données durables et fiables
- **16 exigences** définies pour une durée de 3 ans (prochaine mise à niveau en 2023)
- 16 exigences techniques portant sur plusieurs niveaux de conformité (licences, continuité d'accès, confidentialité/éthique, qualité des données...)
- reflètent les caractéristiques que l'on peut confier à des entrepôts fiables : garantir la fiabilité et la durabilité des dépôts de données
- 160 entrepôts sont aujourd'hui certifiés
- Il existe d'autres certifications, CoreTrustSeal étant la plus (re)connue



— 1. Trouver des données de la recherche : les moteurs de recherche spécialisés

Les moteurs de recherches offrent la possibilité de **chercher des données (de tout type) dans plusieurs entrepôts en même temps.**

Ils centralisent les données provenant à la fois des bases de données scientifiques, des agences gouvernementales et de la presse.

Pensez à utiliser **les filtres** pour gagner en temps et en efficacité.

— 1. Trouver des données de la recherche : les moteurs de recherche spécialisés

Datacite : recense tous les jeux de données ayant bénéficié de son service d'attribution de DOI

Data Search (Elsevier) – bêta : cherche différents types et formats de données dans plusieurs disciplines, à partir de plusieurs sources (entrepôts et articles)

Google Dataset Search – bêta : centralise des milliers de bases de données scientifiques et d'agences gouvernementales

Base Search : collecte et met à disposition des données normalisées et indexées pour plus de 300 000 000 documents (dont + 15M data set)

— 2. Réutiliser les données : conditions de réutilisations

Les **données et les codes sources** déposés dans un entrepôt de données sont destinés à être réutilisés selon les **termes de la licence associée au jeu de données**.

Plusieurs types de licences existent et chacune énonce des **autorisations et des restrictions spécifiques**.

En l'absence de licence spécifique associée à un jeu de données, il faut se reporter aux conditions générales d'utilisation (CGU) du site qui les diffuse.

Point de vigilance : si aucune licence ou mention n'est précisée, cela ne signifie pas pour autant que la réutilisation est libre, car des droits peuvent s'appliquer automatiquement (propriété intellectuelle, protection des données personnelles).

— 2. Réutiliser les données : les différentes licences

Quelle licence choisir pour les données ?

Pour les données ouvertes, la [Loi pour une République Numérique](#) impose de choisir une licence parmi [les licences suivantes](#) :

- Licence ouverte (Etalab)
- ODBL
- CC-BY



Si toutefois, le consortium du projet dans lequel les données ont été créés impose un autre choix, vous devez vous conformer à ce choix.

A noter : les [codes sources](#) sont également concernés.

—2. Réutiliser les données : Etalab



LICENCE OUVERTE
OPEN LICENCE

Dans le cadre de la politique du gouvernement en faveur de l'ouverture des données publiques (Open Data), [Etalab](#) a conçu la « **Licence Ouverte** » libre et gratuite qui :

- Apporte la **sécurité juridique** nécessaire aux producteurs et aux réutilisateurs des données publiques
- **Promeut la réutilisation la plus large** en autorisant la reproduction, la redistribution, l'adaptation et l'exploitation commerciale des données
- S'inscrit dans un contexte international en étant **compatible avec les standards des licences** Open Data développées à l'étranger (ODBL ou CC-BY)

—2. Réutiliser les données : ODBL

La licence Open Database Licence présente les caractéristiques suivantes :

Le réutilisateur de jeux/bases de données mises à disposition sous cette licence peut les :

- Copier, distribuer et utiliser
- Produire de nouvelles créations
- Adapter, modifier et en construire de nouveaux

Sous réserve de :

- **Mentionner impérativement la paternité**
- **Partager aux conditions identiques**



— 2. Réutiliser les données : Licence CC

Les Licences Creative Commons sont :

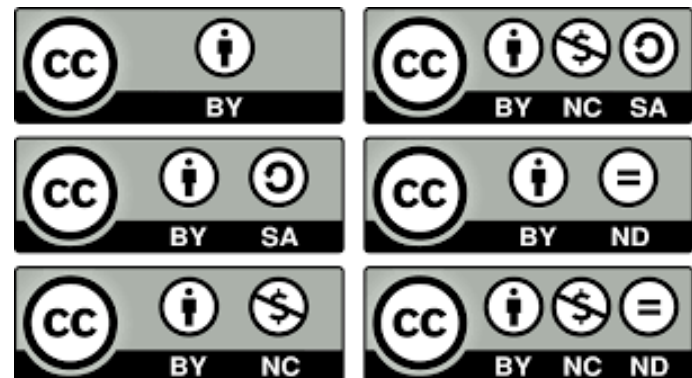


- Les plus utilisées dans le monde
- Utilisables légalement et juridiquement dans tous les pays
- Un **système modulable** constitué de libertés de bases (reproduire, distribuer et communiquer l'œuvre au public), d'une clause permanente (paternité) et de trois clauses optionnelles
- Ne sont pas révocables

Pour en savoir plus :

[Droit d'auteur et licences CC](#)

[Utiliser les licences CC](#)



—3. Déposer les données de la recherche en accord avec les principes FAIR

Quand ?

- A la fin d'un projet de recherche
- Il n'est jamais trop tard : il est possible de le faire de manière rétrospective et curative

Qui ?

- Le responsable du package de données
- Un doctorant en fin de thèse
- Tout le monde est concerné par cette problématique

—3. Déposer les données de la recherche en accord avec les principes FAIR

Est-ce possible ?

- Avez-vous des données personnelles ou sensibles ?
- Quelles sont les conditions de partage et d'ouverture des données avec des partenaires extérieurs ?
- Quel établissement détient les droits de propriété intellectuelle sur les données ?
- Quelle est la législation applicable ?
- Vous pouvez vous référer à l'accord de consortium pour répondre à ces questions.

—3. Déposer les données de la recherche en accord avec les principes FAIR

Quoi ?

- Sélectionner les données à conserver à **moyen terme**
- Se questionner sur la **mort des données**
- Enrichir les métadonnées pour documenter cette sélection
- Publier les données ET les **métadonnées**

— 3. Déposer les données de la recherche en accord avec les principes FAIR

Plusieurs critères sont à prendre en considération pour sélectionner les données à conserver et/ou à détruire :

- La **date et la fréquence** à laquelle faire cette sélection : fin de thèse, de projet ou de contrat / avant de quitter son emploi / à date régulière
- La **durée de conservation** : à définir en fonction des besoins
- L'**obligation administrative** de destruction éventuelle
- La nécessité d'**anonymisation** éventuelle
- Le **format et la nature des données** qui définiront leur lisibilité dans le temps
- Les **supports d'enregistrement, d'utilisation et de stockage** des données
- La **criticité** et donc le niveau de sécurité et d'accessibilité nécessaires pour protéger les données
- Le **coût** de ces supports, des modifications de formats de fichiers, ou bien de l'espace de stockage nécessaire à la conservation des données

—3. Déposer les données de la recherche en accord avec les principes FAIR

Comment ?

- Publier un **data paper**
- Déposer dans un **entrepôt**

— 3. Déposer les données de la recherche en accord avec les principes FAIR : les data papers

Le **data paper** est un **article scientifique sur les données** qui décrit un jeu de données à l'aide d'informations plus précises et détaillées que celles mentionnées dans une publication scientifique traditionnelle.

Pour cela, il insiste sur :

- Les **aspects méthodologiques**
- La **qualité** des données, leurs méthodes de collecte et de traitement
- L'**originalité**, la **portée** et le **potentiel** pour des utilisations futures
- L'**accessibilité** des données (via un fichier attaché au data paper ou un lien pérenne vers un entrepôt où le jeu est déposé et accessible)

— 3. Déposer les données de la recherche en accord avec les principes FAIR : les data papers

Publier un data paper permet de :

- **Valoriser les données**
- **Faciliter leur réutilisation**
- **Leur apporter de la visibilité**
- **Les rendre plus facilement repérables et citables**

A noter : le data paper étant une publication citable, au même titre que tout article scientifique, il **met en valeur ses auteurs** en tant que créateur de données et **permet la traçabilité des citations et des réutilisations**.

—3. Déposer les données de la recherche en accord avec les principes FAIR : le choix de l'entrepôt

Pour publier des données dans un entrepôt, il faut :

- Sélectionner l'entrepôt de données qui **convient le mieux** à votre discipline, votre projet, votre thématique de recherche...
- Si besoin, utiliser un **annuaire** d'entrepôt pour l'identifier
- Attention, il convient de prendre des précautions et de s'assurer de la **qualité de l'entrepôt** avant de réaliser le dépôt

A noter : des entrepôts spécifiques peuvent être suggérés (ou imposés) par le journal dans lequel vous déposez votre article, mais aussi par le financeur, le consortium du projet ou l'institution dans laquelle vous travaillez.

—3. Déposer les données de la recherche dans un entrepôt : les critères de sélection de l'entrepôt

Différents critères sont à prendre en compte :

- Types de données
- Formats de fichiers acceptés
- Volumétrie
- **Métadonnées**
- **Licence**
- **Identifiant pérenne**
- Facilité/aide dépôt
- Gestion des accès (embargo) et des versions
- Liens vers publications et statistiques d'usages
- Localisation du serveur et organisme responsable de l'entrepôt
- Visibilité
- Conservation sécurisée
- Visualisation
- Coût

— 3. Déposer les données de la recherche dans un entrepôt : les identifiants pérennes

Il en existe 2 grands types : les **identifiants objets** pour les productions scientifiques (publications, données, logiciels) et les **identifiants contributeurs** pour les auteurs et les institutions.

Ils favorisent le partage, la réutilisation et l'accessibilité des productions scientifiques sur le long terme et simplifient aussi leur citation.

A titre d'exemple :

- Un **DOI** est un identifiant objet pérenne pour une production scientifique
- Un **ORCID** est un identifiant contributeur pérenne pour un auteur



Source : <https://doranum.fr/identifiants-perennes-pid/>

—4. L'archivage des données et les entrepôts ?

L'archivage consiste à ranger un document dans un lieu où il sera conservé pendant une période plus ou moins longue et d'y associer les moyens pour réutiliser les données : la réutilisation se faisant en ajoutant de l'intelligence à la sauvegarde.

Le contenu des documents archivés n'est pas modifiable. Par contre le contenant (format) des documents archivés peut être modifié (pour éviter l'obsolescence logicielle).

—4. L'archivage des données et les entrepôts ?



Le **Centre Informatique National de l'Enseignement Supérieur (CINES)** propose des moyens exceptionnels à l'ensemble de la communauté scientifique dans ses missions statutaires stratégiques nationales :

- Le calcul numérique intensif
- **L'archivage pérenne de données électroniques**
- L'hébergement de plates-formes informatiques d'envergure nationale
- Attention cela coûte cher !

A noter : les entrepôts de données ne sont donc pas concernés par l'archivage, ils sont à utiliser pour la conservation à moyen terme.

—4. L'archivage des codes sources et les entrepôts ?



Software Heritage est la solution technico-pratique pérenne recommandée pour archiver les **codes sources des logiciels**.

A noter : Github et Gitlab sont à utiliser pour conserver, partager et faire évoluer les codes sources des logiciels à moyen terme.



Cycle de webinaires autour de la Science Ouverte

Les implications pour les chercheurs de la politique SO de l'ANR

05 juillet 2022 de 11h à 12h

Pour revoir des webinaires thématiques, c'est par [ici](#) !

Notamment :

- Introduction à la science ouverte
[consulter le support](#) et [revoir le webinaire](#)
- Principes FAIR et ouverture des données de la recherche
[consulter le support](#) et [revoir le webinaire](#)
- Pourquoi rédiger un plan de gestion de données ?
[consulter le support](#) et [revoir le webinaire](#)
- Comment rédiger son plan de gestion de données avec DMP-Opidor
[consulter le support](#) et [revoir le webinaire](#)

Merci pour votre écoute !

Gardons contact pour la suite :

- Pour des questions autour des publications ou de l'utilisation de HAL : hal-assistance@umontpellier.fr
- Pour des questions autour de la gestion des données de la recherche ou pour toute relecture de PGD : donnees-recherche@umontpellier.fr
- Pour toute autre question : laure.lefrancois@umontpellier.fr

Chargée de mission Appui à la recherche pour l'université

— Webographie récapitulative

Doranum (le site incontournable d'auto-formation pour les données de recherche)

Les entrepôts :

- entrepôts de données
- stockage et l'archivage des données

Les licences :

- Droit d'auteur et licences CC
- Utiliser les licences CC

Les pratiques FAIR :

- pratiques FAIR

Les data papers :

- <https://www.datacc.org/vos-besoins/valoriser-ses-donnees/data-papers/>
- <https://doranum.fr/data-paper-data-journal/>
- <https://mi-gt-donnees.pages.math.unistra.fr/site/FAQ.html#datapaper>