



Traque tes données



Comment trouver des données utiles à sa recherche ?



Différents lieux de diffusion des données

- des « **entrepôts** »
- des **données directement liées aux publications**



I- Les entrepôts de données



I- Les entrepôts de données

I-1 : Définition et présentation



Définition

Service en ligne permettant le dépôt, la description, la conservation, la recherche et la diffusion des jeux de données.

Plus de 2500 entrepôts !!

Intérêt

- Recommandation des financeurs (ANR, Horizon Europe)
- Exigence des éditeurs... (Springer, Elsevier, Taylor and Francis)



Différents **types d'entrepôts** :

Généralistes

- Zenodo (CERN)



- Dryad



- Figshare



- Mendeley data (Elsevier)

- Science Data Bank (ScienceDB)





Différents types d'entrepôts :

Nationaux

Data Archiving and Networked Services (DANS)

Données publiques

Data gouv : plateforme ouverte des données publiques françaises

Legifrance (droit français : lois et jurisprudence!)

Data europe : le portail officiel des données européennes

US data gouv : U.S. Government's open data

UNESCO Institute for Statistics (UIS)



Par **institution/organisme**

Data INRAE

Datasud (IRD)

Data.sciencespo

Harvard dataverse

Appollo (Cambridge University repository)

ESRF Data Portal



Des entrepôts **thématiques**

Science des matériaux

- Materials Cloud Archive
- Crystallography Open Database (COD)
- MassBank

Chimie

- PubChem
- ChEMBL
- ioChem-BD



Sciences de l'environnement

- PANGAEA
- RESIF
- Pôle National de Données de Biodiversité (PNDB)
- International Virtual Observatory Alliance (IVOA)
- Incorporated Research Institutions for Seismology (IRIS)
- Magnetism Information Consortium (MagIC)
- World Data System
- Global Biodiversity Information Facility (GBIF)
- DATA TERRA
- Centre de Données astronomiques de Strasbourg (CDS)



Sciences sociales

- Inter-university Consortium for Political and Social Research (ICPSR)
- Progedo/Quetelnet
- Qualitative Data Repository (QDR)
- UK Data service
- Linguistic Data Consortium
- Open science framework (OSF)
- Nakala (HumaNum)
- Archaeology Data Service



Sciences médicales, biologie

- WHO International Clinical Trials Registry Platform (ICTRP)
- Omics Discovery Index
- wwPDB (Protein Data Bank)
- GenBank
- GEO for genomic datasets
- Uniprot
- Genome Sequence Archive (GSA)
- Global Biodiversity Information Facility (GBIF)



Codes et logiciels

Software Heritage

>>> Lien avec HAL



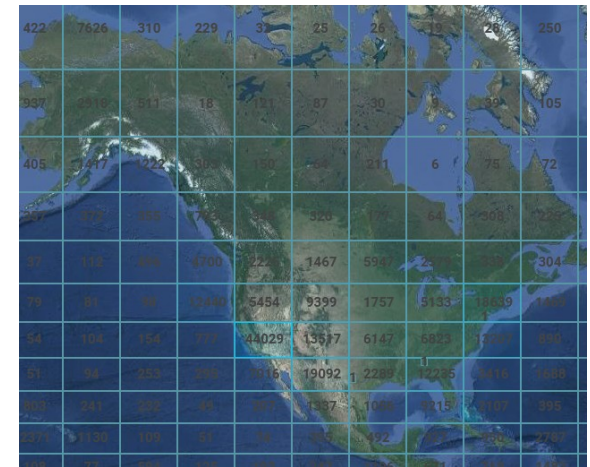
Trouver des données : les entrepôts (12)



L'intérêt des entrepôts : des **fonctionnalités** utiles pour faciliter l'identification et la récupération des données ..

Comment trouver un jeu pertinent ?

Moteur de recherche
cartes



Filtres (sujets, type, format, accessibilité, géographiques/chronologiques, organismes, classifications, etc)

Des outils de recherche (plugin, requêtes API, webservices, paquets R et python, etc..)



- Que vais-je vraiment trouver dans ces données ?

Description détaillée : titre, auteur, mots-clés/taxonomie, résumé, couverture, format, type, taille, versions, organismes/financeurs, objectifs, paramètres, read-me, etc.)

voir par ex :

<https://doi.pangaea.de/10.1594/PANGAEA.842378>

<https://www.icpsr.umich.edu/web/ICPSR/studies/7610>

<https://datadryad.org/stash/dataset/doi:10.5061%2Fdryad.dc1s8h3>

Prévisualisation



Y-a-t-il d'autres ressources intéressantes en lien avec cette donnée ?

- **Données en rapport** voire données utilisées par d'autres (« Users interested in this dataset were also interested in ... »)

- **Publications**

Qu'ai-je le **droit** d'en faire ?

Licences

License (for files):

[CC BY](#) Creative Commons Attribution 4.0 International



Comment vais-je **les récupérer** ?

Modalités d'export métadonnées (RIS, bibteX) et données (zip, etc.)

Comment vais-je **les citer** dans ma thèse

Identifiant pérenne (doi) - **citabilité**

Institute of Meteorology and Geophysics (2013): Climate Data Vent, Ötztal Alps, 1935-2011. University of Innsbruck, PANGAEA, <https://doi.org/10.1594/PANGAEA.806582>

Ohio State University. Center for Human Resource Research. National Longitudinal Surveys of Labor Market Experience, 1966-1992. Inter-university Consortium for Political and Social Research [distributor], 2008-05-21. <https://doi.org/10.3886/ICPSR07610.v1>



Puis-je avoir confiance ?

Modération ?

Certification ?

Durée d'accès ?



Indexation / moissonnage par les moteurs de recherche



I- Les entrepôts

I-2 : Comment trouver un entrepôt pertinent pour sa recherche ?



Répertoire d'entrepôts

Re3data (Registry of Research Data Repositories)

Repository Finder (Datacite)

OAD(Open Access Directory) Data Repositories

OpenDOAR

CoreTrustSeal (liste des entrepôts certifiés)

Cat OPIDoR (entrepôts français)

Dataacc.org (physique, chimie, science de l'environnement)

Fairsharing (sciences de la vie et biomédecine) : entrepôts et bases de données



Name of repository	PANGAEA
Additional name(s)	Data Publisher for Earth and Environmental Science
Repository URL	https://www.pangaea.de/
Subject(s)	Oceanography Geology and Palaeontology Geophysics Geochemistry Mineralogy and Crystallography Biology Atmospheric Science and Oceanography Geosciences (including Geography) Natural Sciences Geology and Palaeontology Geophysics and Geodesy Geochemistry Mineralogy and Crystallography Life Sciences
Description	PANGAEA - Data Publisher for Earth & Environmental Sciences has an almost 30-year history as an open-access library for archiving, publishing, and disseminating georeferenced data from the Earth, environmental, and biodiversity sciences. Originally evolving from a database for sediment cores, it is operated as a joint facility of the Alfred Wegener Institute, Helmholtz Centre for Polar and Marine Research (AWI) and the Center for Marine Environmental Sciences (MARUM) at the University of Bremen. PANGAEA holds a mandate from the World Meteorological Organization (WMO) and is accredited as a World Radiation Monitoring Center (WRMC). It was further accredited as a World Data Center by the International Council for Science (ICS) in 2001 and has been certified with the Core Trust Seal since 2019. The successful cooperation between PANGAEA and the publishing industry along with the correspondent technical implementation enables the cross-referencing of scientific publications and datasets archived as supplements to these publications. PANGAEA is the recommended data repository of numerous international scientific journals.
Contact	https://www.pangaea.de/contact/ info@pangaea.de
Content type(s)	Source code Standard office documents Images Plain text Archived data Audiovisual data
Certificates and Standards	CoreTrustSeal
Keyword(s)	agriculture atmosphere biology biosphere cryosphere earth science ecology environmental science fisheries land surface lithosphere paleontology
Persistent identifier(s) of the repository	FAIRsharing_doi:10.25504/FAIRsharing.6yw6cp
Repository type(s)	disciplinary
Mission statement for designated community	https://www.pangaea.de/about/
Research data repository language(s)	English
Data and/or service provider	data provider

<https://www.re3data.org/repository/r3d100010134>



II- Les données publiées



Les données publiées

- Données intégrées à un article
- « Supplementary materials » (en association avec un article)
- Datapapers (datajournals)



Les **datapapers** (data articles / data descriptors)

Définition :

Articles qui ont pour but de **décrire un ou plusieurs jeux de données**, plutôt que des résultats d'analyse. dans des revues classiques ou dans des revues spécifiques « data journals »

Caractéristiques

- **Données souvent partielles**
- **Reviewing**
- Citation aisée
- Réutilisation des données plus facile que dans les supplementary materials

Pratique :

Les données peuvent être **déposées dans un entrepôt**, recommandé par l'éditeur ou au choix de l'auteur



Un exemple : article dans « open archeology data »

Quelques **listes** :

- Liste du wiki de l'université d'Edinburgh
- Liste de la Forschungsdaten.org
- La base « Où publier » du Cirad (sélectionner data papers)
- La liste du Global Biodiversity Information Facility (Gbif)
- La liste présente sur dataacc' (physique, chimie, sciences de l'environnement, sciences de l'ingénieur)
- Liste élaborée par le projet PREPARDE (univ Leceister)
- Liste de l'université du Michigan
- Titres dans le Web od Science (master journal list ; rechercher « data »)



III - Comment trouver un jeu de données correspondant à sa recherche ?

Quelques outils ...



Des **moteurs de recherche** « données » pour vous aider

- Datacite Search
- Dataset Search (Google)
- DataSearch (Elsevier)
- Data Citation Index (Clarivate), sous abonnement
- Registry of Open Data on AWS (RODA)



Des **moteurs de recherche** « généralistes » pour lier les informations

- Dimensions et Lens (lien avec les publications et brevets)
- Bielefeld Academic Search Engine (BASE) : filtre dataset/software
- OpenAire (Union européenne : liens publications données projets de recherche)
- Isidore (SHS)



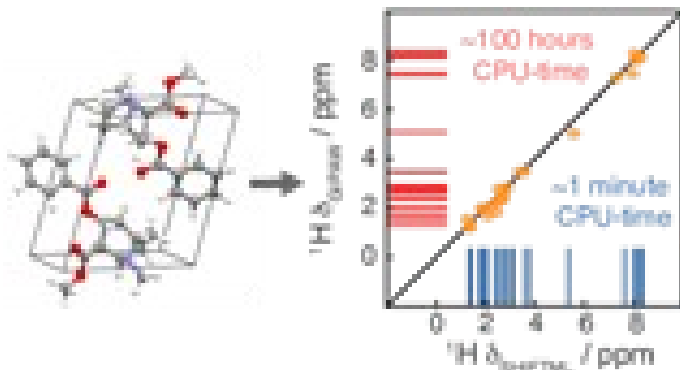
IV- Focus juridique - Qu'ai je le droit de faire avec les données trouvées ?

Trouver des données : le droit (2)

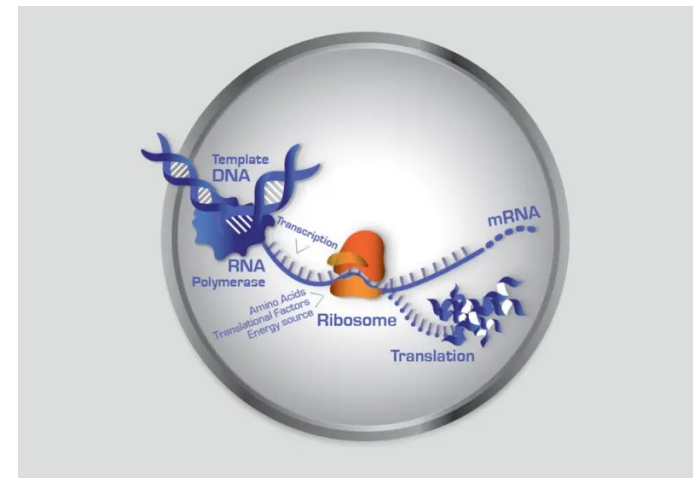


Concerne tous les types de données

- les textes, les écrits
- les images
- les captations audios et vidéos
- les textes
- les modèles de simulation – les codes – les logiciels
- les expressions artistiques (peintures, etc.)



<https://www.materialscloud.org/work/tools/options2022/04/27>



<https://agentmajeur.fr/illustration-scientifique/2022/04/27>

Trouver des données : le droit (3)



LES DROITS MORAUX

Essentiellement le droit de paternité :

Droit d'être reconnu
comme le père (l'auteur)
Droit d'être cité

Droits inaliénables et
imprescriptibles

LES DROITS PATRIMONIAUX

Essentiellement le droit de reproduction :

Droit d'être publié
Droit d'être copié
Droit de diffuser (colloque par ex)

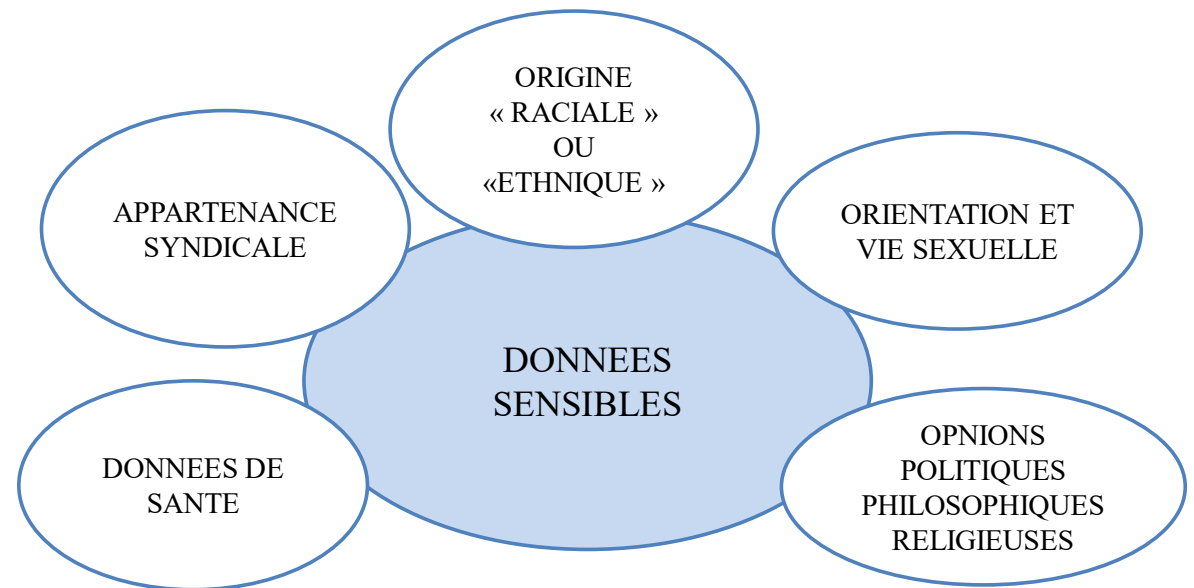
Droits Cessibles
et temporaires
(70 ans)



Une autorisation indispensable :

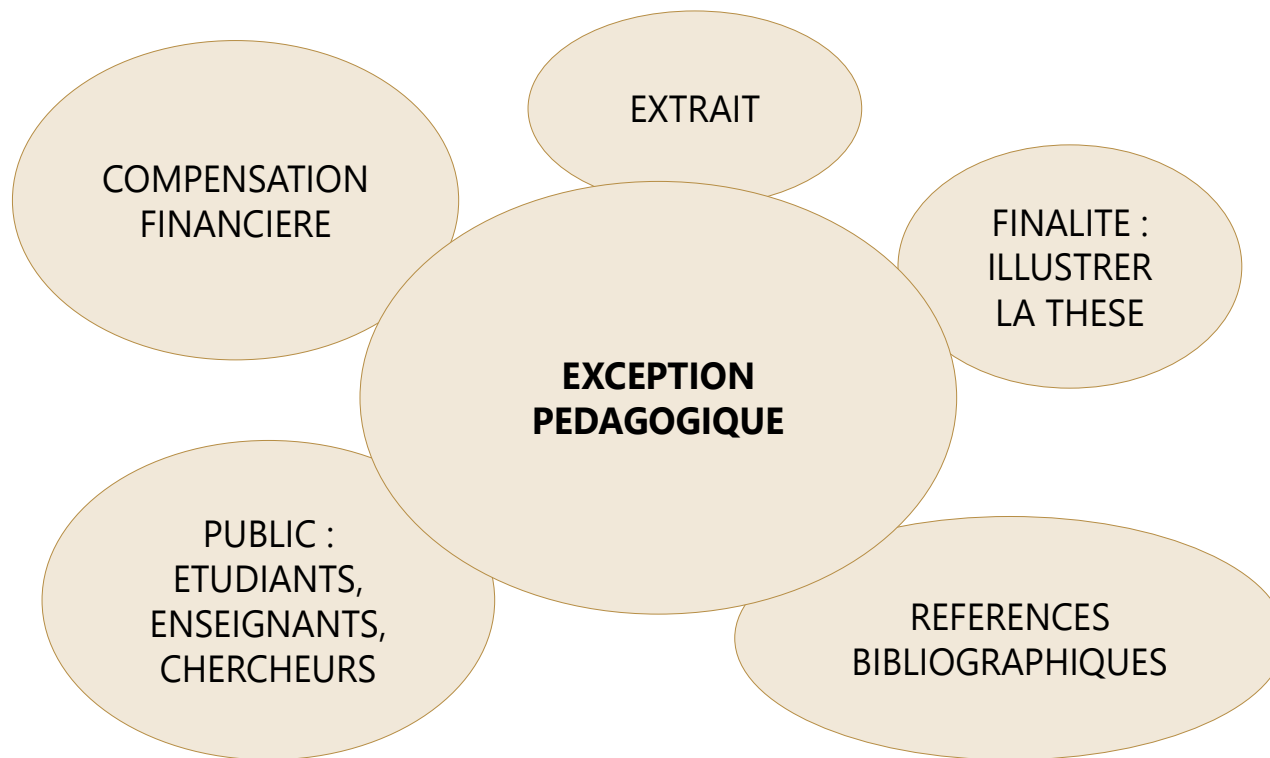
- pour les données personnelles et sensibles
- pour les captations audio et vidéo
- pour l'utilisation de marques ou de brevets

LES DONNEES PERSONNELLES SENSIBLES



Par principe la loi interdit le traitement des données sensibles sous peine de 5 ans d'emprisonnement et 300.000 euros d'amende.

L'EXCEPTION PEDAGOGIQUE



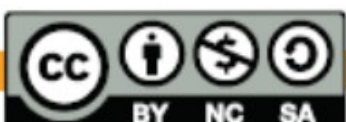
Trouver des données : le droit (6)



Les œuvres écrites	✓ quelques lignes d'un texte d'une page ✓ quelques dizaines de lignes d'un texte de quelques pages ✓ quelques pages d'un ouvrage de plusieurs dizaines de pages
Les images	✓ Pas plus de 20 œuvres par travail de recherche ✓ Une définition limitée à 400/400 pixels et 72 DPI ✓ Impossibilité d'extraire l'œuvre.
Les œuvres audiovisuelles ou cinématographiques	✓ Six minutes au maximum ✓ moins d'un dixième de la durée totale de l'œuvre intégrale ✓ durée totale des extraits moins de 15 % de la durée totale de l'œuvre
Les extraits musicaux	✓ Pas plus de 30 secondes ET ✓ Pas plus de 15 % de la durée totale
Les partitions	✓ Interdiction de reproduction

Au-delà il faut demander l'autorisation à l'auteur ou à ses ayants droits.

Les licences Creative Commons

		Utilisation Partage	Adaptation Modification	Utilisation commerciale	Modification de licence	
TRÈS LIBRE						- Utilisation commerciale autorisée - Modifications ou remix autorisés
						- Utilisation commerciale autorisée - Modifications ou remix autorisés - Les versions dérivées de l'œuvre doivent conserver la licence originale ou compatible
LIBRE						- Utilisation commerciale NON permise - Modifications ou remix autorisés
						- Utilisation commerciale NON permise - Modifications ou remix autorisés - Les versions dérivées de l'œuvre doivent conserver la licence originale ou compatible
NON LIBRE						- Utilisation commerciale autorisée - Modifications ou remix NON permis
						- Utilisation commerciale NON permise - Modifications ou remix NON permis



BY

ATTRIBUTION

Vous pouvez retenir, réutiliser, réviser, remixer et redistribuer.

L'auteur doit être cité



SA

PARTAGE DANS LES MÊMES CONDITIONS

Vous pouvez retenir, réutiliser, réviser, remixer et redistribuer.

Partage sous licence compatible



NC

POUR USAGE NON COMMERCIAL

Vous pouvez retenir, réutiliser, réviser, remixer et redistribuer.

Pour usage non commercial



ND

PAS DE MODIFICATION

Vous pouvez retenir, réutiliser et redistribuer.

Aucune modification permise