Science ouverte
Université Grenoble Alpes

CELLULE DATA
GRENOBLE ALPES

Valorisation des données scientifiques : publication et diffusion / partage

Formation doctorale 31/05/2022

Plan

Pourquoi diffuser ?

Que diffuser ?

Comment diffuser ?
- A quelles conditions d’accès ?
- Licences

Quand diffuser ?

Plan

Où diffuser ? Entrepôts

>>>> critères de sélection

Où diffuser ? Publication

Articles / supplementary materials

Datapapers

Propositions de l’UGA

Aide et ressources

Pourquoi diffuser

Processus de publication :

1 - Article scientifique : principaux résultats d'un travail scientifique

2 - Dépôt des données : valeur ajoutée diffusion des données - reproductibilité

3 - Diffusion thèses : sites institutionnels dédiés - plus de détails scientifiques et techniques

Pourquoi diffuser

Pourquoi ne pas partager ?

- Données **personnelles** et données **sensibles**
  - Par ex, environnementales, santé,
    - >>> Procédures spécifiques (anonymisation…)
  - Données présentant des risques
    - défense nationale
    - sécurité publique, des Etats, établissements
  - Données liées à une zone à régime restrictif (ZRR)
  - Données liées à des secrets professionnels

Pourquoi diffuser

Pourquoi ne pas partager ?

- Données comportant une **valeur économique**
  - Brevet
  - Convention/contrat avec des entreprises
    - En vérifier les termes
    - Discussions nécessaires

- **Publication en cours** de préparation
  - Possibilité de déposer ses données associées et de mettre un embargo avant la diffusion

Pourquoi diffuser ses données ?

Pourquoi partager ?

Des données **précieuses ou uniques**

- Coût de production élevé
- Coût de traitement élevé
- Captation unique :
    - ex données astronomiques…

Augmenter sa visibilité

- Données accessibles et citables indépendamment de l'article
- Lier les données à ses publications
- >>> Augmenter la visibilité de ses recherches,

Pourquoi diffuser ses données ?

Intégrité et ré-utilisabilité

- **Ethique** et intégrité scientifique (voir décret, décembre 2021)
- Garantir la **reproductibilité** des résultats
    - Fiabilité
    - Transparence
- Assurer la **re-utilisabilité** des données
    - Intérêt pour d'autres projets scientifiques

>>>> Favoriser le progrès de la recherche et l'émergence de nouvelles recherches

>>> Mettre en oeuvre les principes **FAIR** pour les codes et les données
Findable, Accessible, Interoperable, Reusable

Pourquoi diffuser ses données ?

Répondre aux exigences des éditeurs:

Préconisation pour l'accès aux données liées (data sharing)
Committee for Publication Ethics (COPE)
Transparency and Openness Promotion (TOP) Guidelines (Centre pour la Science Ouverte)

Quelques exemples
EDP Sciences
Frontiers (Materials and data policies)
Plos One
Elsevier
Springer / Nature
Taylor and Francis

Pourquoi diffuser ses données ?

Répondre aux exigences des financeurs

- **Horizon Europe 2021-2027** :

La «science ouverte» deviendra le mode opératoire d'Horizon Europe. Il exigera donc un accès ouvert aux publications et aux données.»

- ANR (contribuer à l'ouverture des données quand c'est possible)
- NIH Data Management and Sharing Policy

Répondre aux exigences des **Etats et établissements**

- 2e Plan National pour la Science Ouverte (2021-2024)

Axe 2 : Structurer, partager et ouvrir les données de la recherche
Axe 3 Axe 3 : Ouvrir et promouvoir les codes sources produits par la recherche

- CNRS : Plan données de la recherche (nov 2020)
- Charte du CEA pour la science ouverte (2021)

Données de la recherche et publications scientifiques

Les données publiées dans les articles scientifiques représentent seulement la "partie émergée de l'iceberg".

Plus on descend vers la base de la pyramide, plus il est difficile d'établir un lien entre les articles et les données sous-jacentes.

Beaucoup de données restent non diffusées

Schéma adapté de Report on integration of data and publications. Opportunities for Data Exchange (Reilly S. et al., 2011)

Données de la recherche et publications scientifiques dans *Une introduction à la gestion et au partage des données de la recherche,* INIST

Que diffuser ?

- **Analyse des données**
  - **Est-ce que les données brutes suffisent à assurer la reproductibilité ou la réutilisation**
  - **Est-ce que les données ont nécessité un pré-traitement (analyses) ?**
    - coûteux en temps ou en ressources
    - diffusion du processus de pré-traitement / analyses
  - **Est-ce que des outils/codes/logiciels sont nécessaires pour exploiter les données (pour pouvoir utiliser les analyses par exemple)**
    - indiquer les outils nécessaires pour l'exploitation des données
    - diffuser si possible les outils/codes/logiciels liés

Impact environnemental : limiter le volume de données

Que diffuser ?

- **Critère d'utilisabilité**
  - Quelles données répondent aux objectifs du projet ?
  - Quelles données peuvent représenter un intérêt pour la communauté?
- **Statut des données**
  - Est-ce que ces données existent déjà ?
    - Si oui, vérifier la pérennité de la source des données
  - Est-ce que ces données sont uniques ?
    - Si oui, les diffuser
    - Si non, déterminer s'il vaut mieux diffuser les données ou le processus les ayant fournies (par exemple simulation, code source)

Types de données de la recherche

Il existe différents types de données de la recherche qui diffèrent selon la manière dont les données sont produites et selon leur valeur supposée.

Données d'observation

- capturées en temps réel ;
- habituellement uniques et donc impossibles à reproduire ;
  - Exemples : neuroimagerie, photographie astronomique, données d'enquête.

publicdomainpictures.net

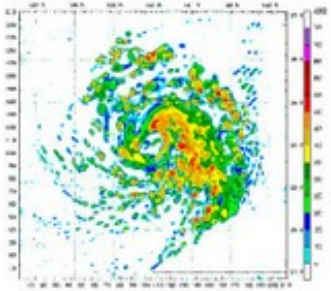
publicdomainpictures.net

Données expérimentales

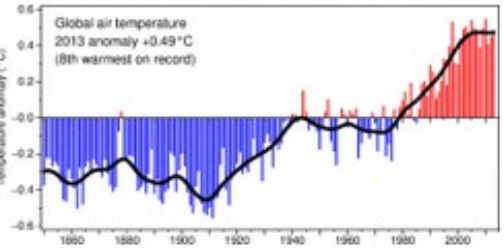
- obtenues à partir d'équipements de laboratoire ;
- souvent reproductibles mais parfois coûteuses ;
  - Exemples : chromatogrammes, puces à ADN.

Données computationnelles ou de simulation

- générées par des modèles informatiques ou de simulation ;
- souvent reproductibles si le modèle est correctement documenté ;
  - Exemples : modèle météorologique, modèle de simulations sismiques, modèle économique.



Typhoon_Mawar_2005_computer_simulation.gif: Atmoz



cru.uea.ac.uk/

Données dérivées ou compilées

- issues du traitement ou de la combinaison de données "brutes" ;
- souvent reproductibles mais coûteuses ;
  - Exemples : fouille de texte, bases de données compilées.

Données de référence

- collection ou accumulation de petits jeux de données qui ont été revus par les pairs, annotés et mis à disposition ;
  - Exemples : GenBank, base de données de cristallographie, collection de lettres ou archive d'images historiques.

publicdomainpictures.net

Données de la recherche et publications scientifiques dans *Une introduction à la gestion et au partage des données de la recherche*, INIST

Que diffuser ?

Plongée dans les spécificités de chacune des communautés

- **Un exemple en physique :**

Les méthodes de publication des données peuvent fortement varier :

- quelques téraoctets avec un accès surveillé pour les données d'astroparticules
- plusieurs pétaoctets pour les données du CERN

Le traitement réservé aux données varie également :

- en physique des hautes énergies, celles-ci sont en général confrontées à des prédictions théoriques avant d’être publiées
- tandis que les données brutes sont diffusées en astronomie

Que diffuser ?

Questions à se poser

- Quel lien avec les publications ?
- Est-il intéressant de diffuser ses données en dehors d'une publication ou sans projet de publication ?

Exemples :

- accès aux données relatives au COVID-19
- données astronomiques où le coût des infrastructures est très élevé

Que diffuser ?

Diffuser des données « négatives » ou « non concluantes » ?

Une diffusion encore trop restreinte

La position de certains **éditeurs**

- Plos One (Collection The Missing Pieces)
- Nature Communication pour les résultats cliniques

La position des **Etats**

- Décret Intégrité scientifique (déc 2021) : « [les établissements] incitent à la publication des résultats de recherche dits négatifs » (art 2)
- Plan National Science Ouverte 2021-2024 : « Réduire le biais de publication, qui est la tendance à ne publier que les études ayant obtenu un résultat positif, au détriment des résultats peu concluants ou négatifs »

- Pour en savoir plus : datacc' + les résultats d'une enquête sur la diffusion des résultats négatifs

Comment diffuser ?

Quel accès ? Plusieurs choix possibles :

- **Libre accès**
  - Immédiat ?
  - Différé ?
    - Accès après un embargo ?
    - Ex : Kéry, Marc, Banderet, Gabriel, Müller, Claudia, Pinaud, David, Savioz, Jérémy, Schmid, Hans, Werner, Stefan, & Monneret, René-Jean. (2021). Spatio-temporal variation in post-recovery dynamics in a large Peregrine Falcon (Falco peregrinus) population in the Jura mountains 2000–2020. Ibis, 164(1), 217–239. https://doi.org/10.5281/zenodo.5862407
- **Accès sur demande**
  - Ex :
  - Ssekuubwa, Enock, van Goor, Wouter, Snoep, Martijn, Riemer, Kars, Wanyama, Fredrick, & Tweheyo, Mnason. (2020). Recovery of seedling community attributes during passive restoration of a tropical moist forest in Uganda [Data set]. In Applied Vegetation Science: Vol. - (-, Number -, p. –). Zenodo. https://doi.org/10.5281/zenodo.4362142
  - Urbanaviciute, Ieva, Bonfiglioli, Luca, & Pagnotta, Mario Augusto. (2022). Raw data: Diversity in root architecture of durum wheat at stem elongation under drought stress (1.0) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.5883299

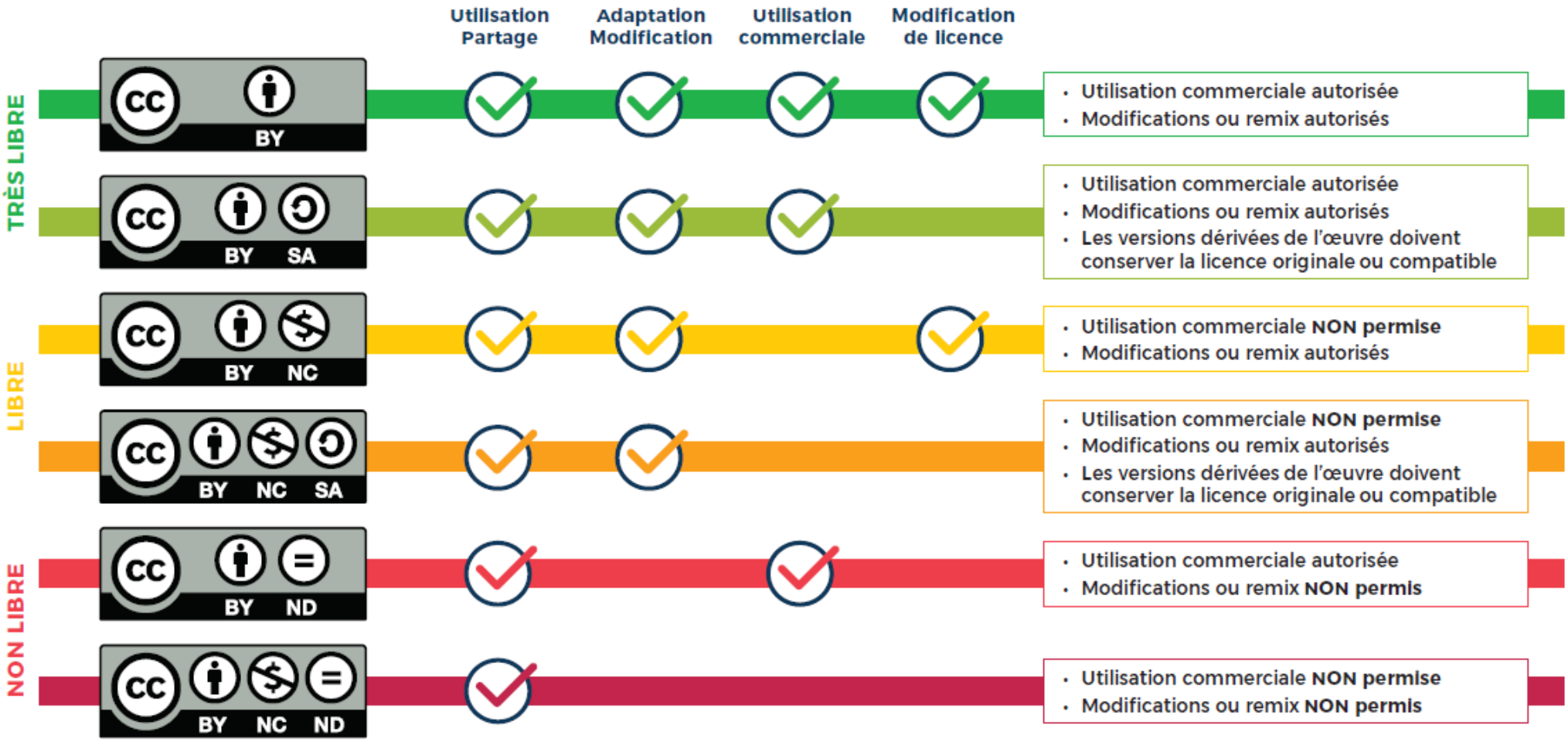
Comment diffuser ? focus licences

Définir l'usage des jeux de données que vous diffusez

Licence : contrat par lequel un titulaire d'un droit de propriété intellectuelle concède en tout ou partie la jouissance de ce droit (droits de reproduction, de représentation, droit d'autoriser les œuvres dérivées…)

- Licences **Creative Commons**
  - BY : Attribution
  - SA : Share Alike (partage à l'identique)
  - NC : Non Commercial
  - ND : Non Derivative (pas de modification)
- Il suffit d'ajouter les logos de chaque famille souhaitée au contenu auquel on veut appliquer la licence

Les licences Creative Commons



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

fabrique REL
RESSOURCES ÉDUCATIVES LIBRES

L'auteur de l'œuvre doit être cité en toutes circonstances pour l'ensemble de ces licences.

Comment diffuser ? focus licences

- Pour les bases de données et les logiciels
  - Pour les bases de données :
    Open Data Commons Open Database License (OdbL)
  - Pour les logiciels :
    - les licences Mozilla, GNU GPL et CeCILL (plus permissives)
    - les licences BSDL, Apache, CeCILL-B et MIT (obligation de réciprocité)

Comment diffuser ? focus licences

- En France, la **Loi pour une république numérique** prévoit une liste de licences (gérée par décret) qui peuvent être utilisées par les administrations pour la réutilisation à titre gratuit de leurs informations publiques : https://www.data.gouv.fr/fr/pages/legal/licences/

  - Etalab 2.0
  - ODC Open Database License (ODbL) version 1.0 (avec obligation de partage à l'identique)

- Compatibilité assurée avec les licences Open Government Licence (OGL, Royaume-Uni), Open Data Commons Attribution (ODC-BY, Open Knowledge Foundation) et Creative Commons Attribution (CC-BY, Creative Commons)

Comment diffuser ?

Focus sur les ZRR et dispositifs de protection du potentiel scientifique et technique de la nation (PPST)

Enjeu : Protection des « savoirs et savoir-faire stratégiques » et technologies sensibles

Protection juridique et administrative

Contrôle des accès physiques et numériques

demande d'autorisation d'accès nécessaire

Les locaux ont le statut de zones protégées : des Zones à Régime Restrictif (ZRR) .

La diffusion de données doit au préalable avoir été expressément autorisée par le responsable de la ZRR

>>> le fonctionnaire de sécurité défense ou d'autres services peut être sollicité.

- Guide de l'agence nationale de sécurité des systèmes d'information (ANSSI) et la direction des affaires internationales, stratégiques et technologiques du SGDSN : Protection numérique du potentiel scientifique et technique de la nation (https://www.ssi.gouv.fr/guide/protection-du-potentiel-scientifique-et-technique-de-la-nation/)

Comment diffuser ?

Définir l'usage des jeux de données que vous diffusez

- **Des ressources pour vous aider :**
  - Contexte juridique de l'Espace chercheurs ENPC et son logigramme dynamique (à qui appartiennent les données / peut-on les diffuser)
  - Arbre **Aide à la décision sur la diffusion des données de recherche (Cirad)**
  - Emilie Cotte, Fanny Sébire. Modèle de logigramme de l'Institut Pasteur relatif aux questions juridiques liées à la réutilisation des données de la recherche. 2022. ⟨pasteur-03587184⟩

- **Ressources DoraNum**
  - Outil d'aide à la décision pour les SHS
  - Autres ressources sur la thématique juridique (RGPD, éthique, doit et open data)

Comment diffuser ?

Définir l'usage des jeux de données que vous diffusez

Choix de licences

Des **outils de sélection de licences** pour les dépôts de données ou de codes :

- Choose an open source licence
- License Selector (codes et données)
- Licentia by Inria

Comment diffuser ?

Définir l'usage des jeux de données que vous diffusez

Des guides

Nicolas Becard, Céline Castets-Renard, Gauthier Chassang, Martin Dantant, Laurence Freyt-Caffin, et al.. Ouverture des données de la recherche. Guide d'analyse du cadre juridique en France. [Rapport de recherche] Comité pour la science ouverte. 2017, 45 p. hal-02791224

Véronique Ginouvès, Isabelle Gras.
La diffusion numérique des données en SHS – Guide des bonnes pratiques éthiques et juridiques, Presses universitaires de Provence, 2018, Digitales, 9791032001790. ⟨hal-01903040⟩

Quand diffuser ?

QUAND DÉCIDE-T-ON DE RENDRE SES DONNÉES PUBLIQUES ?

Il n'y a pas de règles, le mieux est d'ouvrir les données le plus tôt possible

- **Avantages** : vous êtes le premier à produire de nouvelles données
- **Inconvénients** : de nouvelles expériences peuvent confirmer ou non la qualité de vos données

- Un embargo peut aussi être appliqué afin de permettre un délai d'exploitation des données
- Les données sont souvent publiées au moment de la publication des résultats
- Diffuser ses données peut être une justification pour les financeurs

Où diffuser ?

Quel mode de diffusion ?

- Dépôt dans un entrepôt de données

- Publication données intégrées dans un article classique
- Publication de supplementary materials
- Publication d’un data paper

Où diffuser ? Diffuser via un entrepôt

Entrepôt de données :
Service en ligne permettant le **dépôt, la description, la conservation, la recherche et la diffusion des jeux de données** en vue de leur réutilisation.

A ne pas confondre avec des plateformes de stockage ou d'archivage.

Il existe des milliers d'entrepôts !

Différents types d'entrepôts de données

Où diffuser ? Diffuser via un entrepôt

Différents **types d'entrepôts** :

Généralistes

Zenodo (CERN)
Dryad
Figshare
Mendeley data (Elsevier)
Science Data Bank (ScienceDB)

Où diffuser ? Diffuser via un entrepôt

Nationaux

Data Archiving and Networked Services (DANS)
Recherche Data Gouv ! (France, ouverture juillet 2022)

Par **institution/organisme**

Data INRAE
Datasud (IRD)
Data.sciencespo

ESRF Data Portal

Où diffuser ? Diffuser via un entrepôt

Des entrepôts **thématiques**

Sciences de l'environnement

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

Où diffuser ? Diffuser via un entrepôt

Science des matériaux

Materials Cloud Archive
Crystallography Open Database (COD)
MassBank

Chimie

PubChem
ChEMBL
ioChem-BD

Où diffuser ? Diffuser via un entrepôt

Sciences sociales

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

Où diffuser ? Diffuser via un entrepôt

Sciences médicales, biologie

WHO InternationalClinicalTrialsRegistryPlatform (ICTRP)
EMBL's European Bioinformatics Institute (EMBL-EBI)
National Center for Biotechnology Information (NCBI)
Omics Data index
wwPDB (Protein Data Bank)
GenBank
GEO for genomic datasets
Uniprot
Genome Sequence Archive (GSA)
Institut Français de Bioinformatique (IFB)

Codes – logiciels

Software Heritage

Où diffuser ? via un entrepôt

Quel entrepôt choisir ?

- En premier lieu s'il en existe, dans un entrepôt disciplinaire
  - Pratique communautaire
- Recommandations du financeur, de l'éditeur
- Recommandations du projet de recherche, des partenaires
- Recommandations de l'établissement ou l'organisme de rattachement

Où diffuser ? via un entrepôt

Les critères de choix

- Statut et politique de l'entrepôt
  - Disciplinaire ? Généraliste ?
  - Le statut public / privé de l'entrepôt?
  - Lieu d'hébergement du serveur?
  - Certification ? Reconnu ?
  - Modération des dépôts et quel type de modération ?
  - Modèle économique de l'entrepôt (Coût du dépôt?)
  - Origine de l'entrepôt ?
    - Qui est responsable de l'entrepôt ?
  - La préservation sur le long terme des données?

Où diffuser ? via un entrepôt

Les critères de choix (suite)

- Modalités de dépôt
  - Types de données acceptés
  - Type de formats acceptés
  - Identifiant pérenne (doi)
  - Qualité de la description (qualité des métadonnées)
    - Standards ?
  - Gestion des versions
  - Lien avec la publication
  - Volume accepté
  - Type d’accès possibles aux données
    - Embargo / restriction de l’accès
  - Licences

Où diffuser ? via un entrepôt

Les critères de choix (suite)

Autres services

Simplicité d'utilisation
- pour le déposant : facilité du dépôt (formulaire)
- pour les utilisateurs : facilité de la recherche (moteur de recherche, filtres, API, cartes ...)

Aide au dépôt/adresse support

Moissonnage vers d'autres catalogues / entrepôts
Existence d'outils d'outils d'exploitation ou de visualisation

Existence de statistiques d'utilisation, de consultation, de téléchargements

Où diffuser ? via un entrepôt

Pour vous aider à choisir un entrepôt :

Utilisation d'un annuaire pour identifier un entrepôt dans sa discipline :

- re3data (Registry of Research Data Repositories)
- OAD (Open Acces Directory/Data repositories)
- FAIRsharing (sciences de la vie et biomédecine)
- OpenDOAR
- ROAR (Registry of Open Access Repositories)
- CoreTrustSeal (entrepôts certifiés)

etc.

Où diffuser ? via un entrepôt

Des outils :

Trouver un entrepôt de données (université de Bordeaux) - généraliste

CAT OPIDOR – catalogue des services et entrepôts pour les données de la recherche. (Inist-CNRS)

Repository Finder (DataCite)

Data Repository Finder (Université de Utrecht)

How to find a trustworthy repository for your data (OpenAIRE)

Sansone, Susanna-Assunta, McQuilton, Peter, Cousijn, Helena, Cannon, Matthew, Chan, Wei Mun et al. (2020). Data Repository Selection: Criteria That Matter. Zenodo. https://doi.org/10.5281/zenodo.4084763

Où diffuser ? via un entrepôt

Des outils (disciplinaires) :

DATACC’ – dispositif d'accompagnement à la gestion des données de la recherche en physique et en chimie. (UGA et Lyon 1)

La liste d’entrepôts dans le domaine biomédical (CeRIS)

Enabling FAIR Data Community, Ruth Duerr, Danie Kinkade, Michael Witt, & Lynn Yarmey. (2018). Data Repository Selection Decision Tree for Researchers in the Earth, Space, and Environmental Sciences. Zenodo. https://doi.org/10.5281/zenodo.1475430

Scientific Data Sharing – (National Institutes of Health)

Où diffuser ? via un entrepôt

Etapes pour le dépôt de ses données

- Choix des données pertinentes (brutes, traitées, analysées)
  - Avec un point de vigilance concernant les données personnelles ou confidentielles
  - En s'assurant que les volumes sont en adéquation avec ce que permet l'entrepôt choisi
  - En veillant à mettre les éléments nécessaires pour les utiliser (codes, logiciels, etc.)

- Choix des formats (ouverts)

- Choix des éléments de description : métadonnées générales et disciplinaires, Readme ...

- Organisation et nommage des fichiers

- Choix de la licence et des modalités d'accès (ouvert, restreint, avec embargo)

>>> respecter les principes FAIR (Findable, Accessible, Interoperable, Reusable)

Où diffuser ? via un entrepôt

DOREL

https://dorel.univ-lorraine.fr/dataverse/univ-lorraine

Qu’est-ce que DOREL ?

- DOREL est l’entrepôt des données de l’Université de Lorraine
- DOREL utilise la technologie Dataverse, un logiciel libre de gestion d’entrepôt de données de la recherche.
- Dataverse sera utilisé pour Recherche Data Gouv

Où diffuser ? via un entrepôt

Pre-requis

S'être créé un compte et posséder des droits sur la collection Dataverse de son laboratoire (en faire la demande via l'onglet « support » en cas de besoin)

Avoir un ensemble de fichiers prêts (organisation logique, nommage clair et sans accents), présence d'un README (fichier texte permettant d'expliquer le contenu du jeu de données et d'éventuelles consignes d'utilisation)

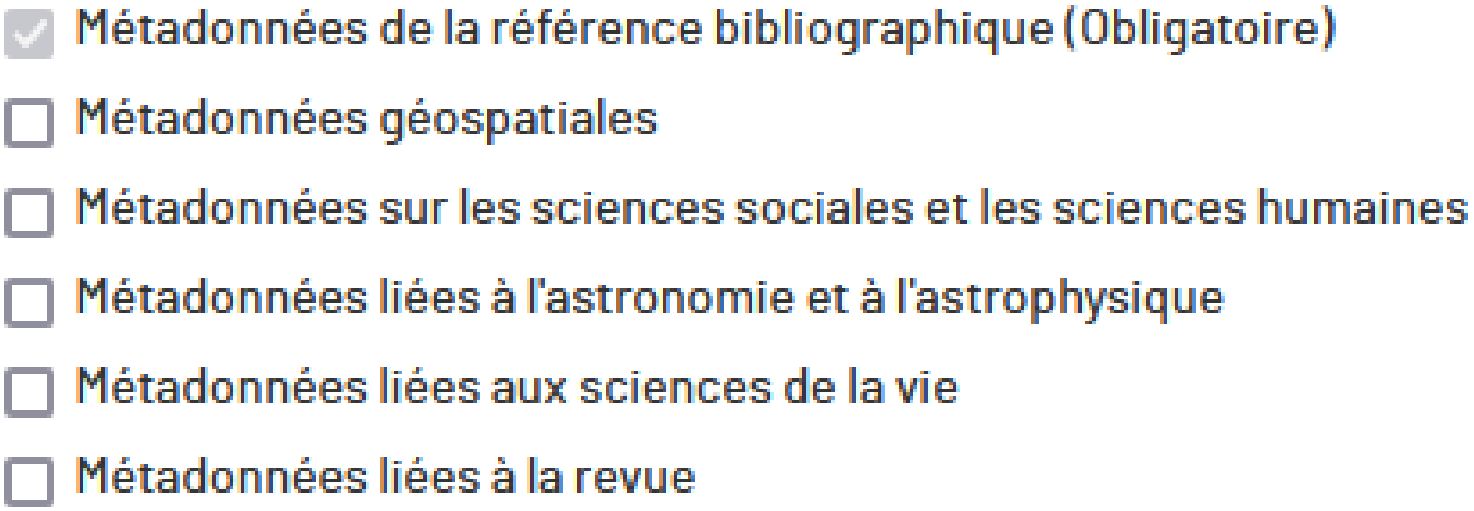
Optionnel : Avoir à portée de main la référence de la publication à laquelle lier le jeu de données

Où diffuser ? via un entrepôt

Possibilité, selon les disciplines, d'adapter le schéma de description des métadonnées

- géographique
- sciences humaines et sociales
- astronomie

Etc.



Où diffuser ? via un entrepôt

Utilisation de Loterre (https://www.loterre.fr)

Loterre (Linked open terminology resources) « est une plateforme d'exposition et de partage de terminologies scientifiques multidisciplinaires et multilingues, conforme aux standards du web des données ouvertes et liées (LOD) ainsi qu'aux principes FAIR. »

Loterre vous permet de trouver des mots-clefs standardisés pour décrire vos données, issus de plusieurs vocabulaires de référence. Son utilisation est recommandée.

⇨ NB : A la date de rédaction de ce guide (automne 2021), Loterre souffre toujours d'un problème sur le champ de recherche en page d'accueil. Il est conseillé de cliquer sur « Explorer » puis « Naviguer » ou se rendre directement à la page https://www.loterre.fr/skosmos/fr/.

Où diffuser ? via un entrepôt

Discipline	Vocabulaire suggéré (champ « Vocabulary »	URL
Multidisciplinaire	UNESCO	http://vocabularies.unesco.org/browser/thesaurus/en/
Sciences physiques	PhySH	https://physh.aps.org/browse?facetIds=Research%2520Areas
Archéologie	PACTOLS	https://masa.hypotheses.org/pactols
Sciences économiques	JEL	https://www.aeaweb.org/jel/guide/jel.php
Environnement	GEMET	https://www.lexicool.com/gemet-multilingual-environment-thesaurus.asp?IL=1
Agriculture	AGROVOC	https://agrovoc.fao.org/browse/agrovoc/en/
Sciences de la vie et santé	MeSH	https://www.loterre.fr/skosmos/JVR/fr/
Psychologie	Thésaurus APA	

Où diffuser ? via une publication

3 manières de publier des données

- Les inclure dans un article (données intégrées ou embedded data)
- Les assembler en annexe dans un matériel supplémentaire (« supplementary materials »)
- Les publier dans un data paper (data article, data descriptor).

Où diffuser ? via une publication

Avantages/inconvénients

Mode de publication	Recherche et citabilité	Paternité et crédits auteurs	Volumétrie	Réutilisabilité
Données intégrées	★★☆☆	★★★★	★☆☆☆	★☆☆☆
Matériel supplémentaire	★★☆☆	★★★★	★★☆☆	★☆☆☆
Data paper	★★★★	★★★★	★★★★	★★★☆

DoRANum. Données de la recherche : apprentissage numérique [En ligne].France : DoRANum; 2017. Comment publier des données de recherche [modifié le 28 mai 2018 ; consulté le 17 septembre 2018]. Disponible : https://doranum.fr/data-paper-data-journal/comment-publier-donnees-recherche/

Où diffuser ? via une publication

Inclure les données directement dans son article

Avantages :

- Intégration maximale dans l'article
- Paternité des données et citabilité

Limites :

- Données souvent très partielles
- Données non disponibles en dehors de l'article
- Données difficilement réutilisables
- Accès de l'article problématique (sauf si diffusion en open access)

Recommandation :

Déposez dans HAL le post-print de la publication

Où diffuser ? via une publication

Mettre les données en annexe de l'article (**matériel supplémentaire**)

Pratique de plus en plus recommandée par les éditeurs pour la validation scientifique

Avantages :
Format des données libéré des contraintes de rédaction de l'article;
- Paternité des données / crédits aux auteurs.

Limites :
- Taille souvent limitée
- Forme peu ou pas réutilisable.
- Peu de standardisation pour la description et le signalement des matériels supplémentaires
- Identification des données indépendamment de l'article possible mais rare ;
- Données difficiles à trouver indépendamment de l'article
- Accès de l'article problématique (sauf si diffusion en open access)

Recommandation
Déposez dans HAL le post-print de la publication

Où diffuser ? via une publication

Publier les données dans **un data paper**

Avantages :
- Paternité des données / crédits aux auteurs ;
- Citation aisée
- Réutilisation des données facilitée ;
- Données normalisées, standardisées, conservées de façon pérenne ;
- Pas de restriction en volume ;
- Liens vers les données déposées réciproques et sécurisés.

Limites :
- Interrogation possible sur la qualité du peer-review
- Interrogation sur l'audience
- Accès de l'article problématique (sauf si diffusion en open access)

Recommandation :
Déposez dans HAL le post print de la publication

Où diffuser ? via une publication

Focus sur les datapapers

Définition :
Les data papers (data articles / data descriptors) sont des articles qui ont pour but de décrire un ou plusieurs jeux de données, plutôt que des résultats d'analyse.
Les data papers peuvent paraître dans des revues classiques ou dans des revues spécifiques « data journals ».

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

Où diffuser ? via une publication

Focus sur les datapapers

Enjeu :
Le datapaper valorise les données en exposant leur potentiel pour des utilisations et projets futurs.

Il facilite la réutilisation des données en mettant en évidence la qualité des données et des procédures, ainsi que la rigueur scientifique de l'étude.

Il apporte de la visibilité aux données, les rend plus facilement repérables et citables par d'autres études.

Le data paper est une publication citable, au même titre que tout article scientifique publié.

Le datapaper est examiné par les pairs

Il permet la traçabilité des citations et des réutilisations.

Où diffuser ? via une publication

Focus sur les datapapers

Quelques exemples de datajournals :
- Data Science Journal – codata
- Scientific Data (Nature)
- Journal of Open Humanities Data(JOHD)
- Data In Brief (Elsevier)

Quelques listes :
- La base « Où publier » du Cirad (sélectionner comme type d'articles data papers)
  La liste du Global Biodiversity Information Facility (Gbif)
- Data journals (Forschungsdaten.org)
- A Growing List of Data Journals, Katherine Akers
- Dedieu, L. 2022. Publier un Data paper, en 5 points. Montpellier (FRA) : CIRAD, 5 p. https://doi.org/10.18167/coopist/0057 (liste commentée)
- Datacc' Publier un data paper : où et comment ? :

- Publier ses codes et logiciels : liste de Software Sustainbility Institute (2021)

- Possible dans de nombreuses revues « classiques » (PlosOne, CyberGeo ...)

Où diffuser ? via une publication

Focus sur les datapapers

<u>Un exemple de plan</u> :
Eléments **communs** avec les autres types d'articles

- Titre,
- Auteur(s)
- Résumé
- Mots clefs,
- Introduction (contexte dans lequel les données ont été obtenues, objectif de recherche)
- Remerciements, sources de financement
- Références,
- Discussions.

Eléments **spécifiques** : Tous types d'informations permettant d'interpréter, de réutiliser les données et de reproduire l'étude

- Le **type de données** (image, tableau, graphe…),
- La couverture géographique et temporelle,
- Les **conditions d'accès et de réutilisation** (licence),
- La taxonomie,
- **L'intérêt** du jeu de données, ce qu'il apporte de nouveau,
- **Les méthodes** utilisées permettant un **contrôle qualité**.

Source : DoraNum, le contenu d'un datapaper, 2018

Où diffuser ? via une publication

Un autre exemple de plan (cf modèle d'articles de données de l'IFSTTAR)

Titre, Auteur, Affiliation ; Data's citation : (telle qu'elle apparait dans l'entrepôt)

1. DATA PRESENTATION

1.1. GENERAL INFORMATION
Petite introduction d'une ou deux phrases présentant de façon très simple, de quels types de données il s'agit (images, cartes, sons, données de capteurs, algorithmes etc.) et sur quoi portent les données.

1.2. DATA FILES
Explication plus approfondie des fichiers de données, une description de l'arborescence des fichiers, des formats

1.3. STRUCTURE OF DATA
Informations relatives aux données (structure des données dans chaque fichier, les mesures et leurs unités) etc.

1.4. VALUE OF THE DATA
Indiquer dans cette partie à quoi servent les données, quelle est leur originalité

2. METHOD
2.1. PROJECT
Pour chaque type de données, expliquer pourquoi vous avez eu besoin de ces données, l'objectif visé dans le projet.

2.2. EXPERIMENTATION
Pour chaque type de données, expliquer les méthodes qui ont permis de les obtenir

2.3. MATERIALS
Pour chaque type de données, expliquer le matériel utilisé, le calibrage de ces outils etc.

Où diffuser ? via une publication

Focus sur les datapapers

Dans tous les cas, se référer
- aux guides/instructions des éditeurs à destination des auteurs
- aux templates mis à disposition par les éditeurs

Exemples :
- Template for data descriptor pour *Scientific Data* (Nature Publishing Group - Overleaf)
– Le template de Data in Brief
- Journal of Open Humanities Data (JOHD) data paper template (open humanities data) Journal of Open Humanities Data (JOHD)
– L'outil arpha Writing Tool.
- L'outil mis à disposition sur DataInrae

Où diffuser ? via une publication

Les **critères d'évaluation** des datapapers
(Doranum, les critères d'évaluation des data papers (2017), DOI : 10.13143/dk71-w757)

- « **Importance et originalité** des données ;
- Potentiel et **valeur de réutilisation** des données ;
- **Qualité et fiabilité des données** : structure de la base de données, organisation logique des données, intégrité des données (vérification des erreurs potentielles) ;
- **Accès aux données** : point important. L'auteur doit s'assurer que ses données sont toujours accessibles dans l'entrepôt. Si les données sont retirées de l'entrepôt, la rétractation du data paper pourra être décidée par l'éditeur ;
- **Qualité et rigueur de la méthode** de collecte des données : méthode appropriée, actuelle, suffisamment claire pour permettre la reproductibilité ;
- **Choix des métadonnées** descriptives et formats: présentation, complétude, degré de précision, etc. ;
- Autres critères « classiques » : qualité générale du manuscrit, citations appropriées, respect des instructions, etc. »

Exemple de critères :
Geoscience data journal

Où diffuser ? via une publication

Focus sur les articles executables

Articles exécutables (Executable Research Article ou Executable paper) : logiciel dynamique qui combine du texte, des données brutes et le code utilisé pour l'analyse. Possibilité de reproduire chaque étape franchie. Outils de visualisation associés

- Guide et exemples (CodaLab Worksheets Documentation)
- Where can I publish executable papers and notebooks? By Steven R Brandt. Software Sustainability Institute
- What is an executable paper ?Jana Lasser (18/06/2020), blog Sozialwissenschaftliche Methodenberatung
- Collection sur Elife

voir :

- Lasser, J. Creating an executable paper is a journey through Open Science. Commun Phys 3, 143 (2020). https://doi.org/10.1038/s42005-020-00403-4
È Michael Nauge. Executable papers : vers une science plus ouverte, transparente et réutilisable. Progedo : Semaine DATA SHS, Dec 2020, Poitiers, France. ⟨hal-03128918⟩

Où diffuser ? via une publication

5 exemples de data papers

- Colgan, W., Wansing, A., Mankoff, K., Lösing, M., Hopper, J., Louden, K., Ebbing, J., Christiansen, F. G., Ingeman-Nielsen, T., Liljedahl, L. C., MacGregor, J. A., Hjartarson, Á., Bernstein, S., Karlsson, N. B., Fuchs, S., Hartikainen, J., Liakka, J., Fausto, R. S., Dahl-Jensen, D., Bjørk, A., Naslund, J.-O., Mørk, F., Martos, Y., Balling, N., Funck, T., Kjeldsen, K. K., Petersen, D., Gregersen, U., Dam, G., Nielsen, T., Khan, S. A., and Løkkegaard, A.: Greenland Geothermal Heat Flow Database and Map (Version 1), Earth Syst. Sci. Data, 14, 2209–2238, https://doi.org/10.5194/essd-14-2209-2022, 2022
- Hosseini, K., Beelen, K., Colavizza, G., & Ardanuy, M. C. (2021). Neural Language Models for Nineteenth-Century English. Journal of Open Humanities Data, 7, 22. DOI: http://doi.org/10.5334/johd.48
- Kovylyaeva, A., Astapov, I., Dmitrieva, A., Borog, V., Osetrova, N., & Yashin, I. (2020). Experimental Data of Muon Hodoscope URAGAN for Investigations of Geoffective Processes in the Heliosphere. Data Science Journal, 19(1), 11. DOI: http://doi.org/10.5334/dsj-2020-011
- Testolini, V. (2021). Data from “Ceramic Technology and Cultural Change in Sicily from the 6th to the 11th Century AD.” PhD Thesis. Journal of Open Archaeology Data, 9, 11. DOI: http://doi.org/10.5334/joad.77
- Selvam, R. M. et al. (2015). Data set for the mass spectrometry based exoproteome analysis of Aspergillus flavus isolates. Data in brief, 2, 42-47. http://dx.doi.org/10.1016/j.dib.2014.12.001

Où diffuser ? via une publication

Focus sur les datapapers

Comment choisir ?

S'informer sur :

- La revue, sa notoriété, son importance dans la communauté
- Son modèle économique (Accès ouvert ? Accès ouvert via APC ?)
- Ses contraintes juridiques (Cession de droit exclusive? Possibilité de mettre des licences CC)
- Ses modalités de diffusion (Dépôt dans un entrepôt possible?)
- Son processus éditorial (Peer reviewing, délais de publication)

Vérifier qu'il reste possible de publier un article de recherche en lien avec les données

Voir Dedieu, L. 2022. Publier un Data paper, en 5 points. Montpellier (FRA) : CIRAD, 5 p. https://doi.org/10.18167/coopist/0057

Où diffuser ? via une publication

Ressources

CNRS, 2021. Guide de bonnes pratiques sur la gestion des données de recherche. Publier un Datapaper pour valoriser et expliciter les données.

INRAE, Pubier un datapaper, r
voir la FAQ
le séminaire : "Les data papers. Une nouvelle forme de valorisation des données scientifiques"

DoRANum, Data papers et Data journals
DoRANum, 2020. Webinaire Data paper - Une incitation à la qualification et à la réutilisation des jeux de données.

Où diffuser ? via une publication

Ressources

Publier un datapaper (CIRAD)

Data papers INRAE (bibliothèque partagée Zotero)

Data Papers : quand ? Comment ? Pourquoi ?
GT Science ouverte Couperin

Software Sustainability Institute, 2021.
In which journals should I publish my software?

Où diffuser ? via une publication

Rappel juridique

Déposer dans une archive ouverte un article, un article avec supplementary materials, un datapaper

Rappel :

- Ce que permet la **loi pour une République Numérique, art 30** (2016)
  - Dépôt possible de l'article accepté pour publication pour toute recherche financée au moins à 50 % sur des fonds publics (salaire, équipement, etc)
  - Embargo max de 6 mois
  - Quelque soit le contrat signé avec l'éditeur
  - Voir la FAQ

- Ce qu'exigent les **financeurs** (Coalition S, Horizon Europe, ANR)
  - Diffusion en libre accès immédiat sous licence CC-by des publications scientifiques financées

Ou déposer ses données quand on est dans un laboratoire UGA ?

Où diffuser ses données à l'UGA s'il n'existe pas d'entrepôt thématique pertinent ?

La **plateforme nationale** Recherche Data Gouv

Objectif :
Accompagner les chercheur autour des données.
Proposer une solution de dépôt dans un entrepôt national de confiance

3 volets :

- **Entrepôt** pour rechercher, déposer et publier des **données**
- **Catalogue** pour signaler des données déposées dans des entrepôts externes **(à venir)**
- **Accompagnement** : ateliers de la donnée, centres de ressources (INIST) et de référence thématiques

Ou déposer ses données quand on est dans un laboratoire UGA ?

Choix de l'UGA : proposer un **entrepôt institutionnel** pour répondre aux besoins des scientifiques qui n'ont pas de solution disciplinaire

>>>>> Collection UGA dans la plateforme nationale **Recherche Data Gouv** dès son démarrage en juillet 2022 (à utiliser!!!!)

organisation en collections par laboratoires, projets de recherche, etc

Tous les types de données sont acceptés

Taille max des jeux de données : 50 Go par fichier. Via l'interface web, téléversement possible jusqu'à 1 000 fichiers, au-delà nécessité de passer par l'API RDG (qui est celle de Dataverse). Possible de déposer un fichier zip (contenant au plus 1 000 fichiers) comme jeu de données.

Accès restreint possible pour certains ou tous les fichiers d'un jeu de données.

Possible de créer une **url privée** pour un jeu de données non publié, par exemple pour de la relecture par les pairs

A noter : **privilégier un entrepôt thématique** reconnu par sa communauté

Ou déposer ses données quand on est dans un laboratoire UGA ?

- Services associés côté UGA
  - Création de collections propres aux laboratoires, aux projets de recherche
  - Aide au dépôt (description des données, diffusion)
  - Modération technique locale (dans les labos ou par la cellule data Grenoble Alpes
  - Formations….

- Services associés côté national
  - Centres de reférence thématiques (DataTerra, CDS, PNDB, IFB, Progedo, HumaNum)
  - Centres de ressources (formations, etc.)
  - Recommandations d'entrepôts disciplinaires
  - Guides, FAQ, outils, webinaires …

Aide et ressources

Une adresse mail :
sos-data[at]univ-grenoble-alpes.fr

La cellule data Grenoble Alpes : structure opérationnelle pour répondre concrètement à toutes les demandes des communautés scientifiques de Grenoble sur les données.

Point d'entrée unique :

Aide à la diffusion des données et des codes
Aide à la description des données
Lien publications/données/codes
Aide juridique
Diffusion des bonnes pratiques

Aide et ressources

Un **portail HAL Université Grenoble Alpes** avec une cellule d'accompagnement

Une **cellule Hal UGA** (10 personnes)
Un réseau de correspondants Hal dans les labos

Exemples de **services**
- Aide à la publication en open access
- Aide au dépôt dans hal
- Aide à la gestion de ses identifiants numériques pour la recherche
>>> voir les ateliers Créez votre identifiant IdHal
- Aide à l'affichage et à l'exportation de ses listes de publications, création de CV
- Accompagnement juridique
- Formations, ateliers,
- Veille
- Fiches pratiques
- une **adresse support : hal-support@univ-grenoble-alpes.fr**

>>>> Lien avec la cellule data Grenoble Alpes

Aide et ressources

Equipe de la BU : traitement des thèses et accompagnement des doctorants

Rôle :
- Signalement de la thèse sur thèses.fr
>>> Dart Europe
- Diffusion selon le choix du doctorant (Hal-Thèses en ligne - TEL, intranet)
- Archivage numérique pérenne

Quelques exemples de services :
- Atelier « Je dépose ma thèse »
- Aide juridique
- Conseils sur la diffusion de la thèse
- formations/ateliers : « Entrer dans la communauté des chercheurs »

Service des thèses - une adresse :
bu-theses@univ-grenoble-alpes.fr

Aide et ressources

- Le site science ouverte de l'UGA

- DoraNum
  - Dépôt et entrepôts
  - Data papers et data journals
- INIST
  Une introduction à la gestion et au partage des données de la recherche
  Cours "Comprendre la science ouverte" (connexion anonyme)
- Site du CIRAD, Coopist, Gérer des données
- Guide de bonnes pratiques sur la gestion des données de la recherche, CNRS, chap 7 : Publier et diffuser
- Recherche Data Gouv