HAL open science

Introduction aux données de la recherche

Annabelle Filatre, Hensens Hanka

▶ **To cite this version:**

Annabelle Filatre, Hensens Hanka. Introduction aux données de la recherche. Doctorat. Montpellier, France. 2019, pp.48. hal-04448418

HAL Id: hal-04448418

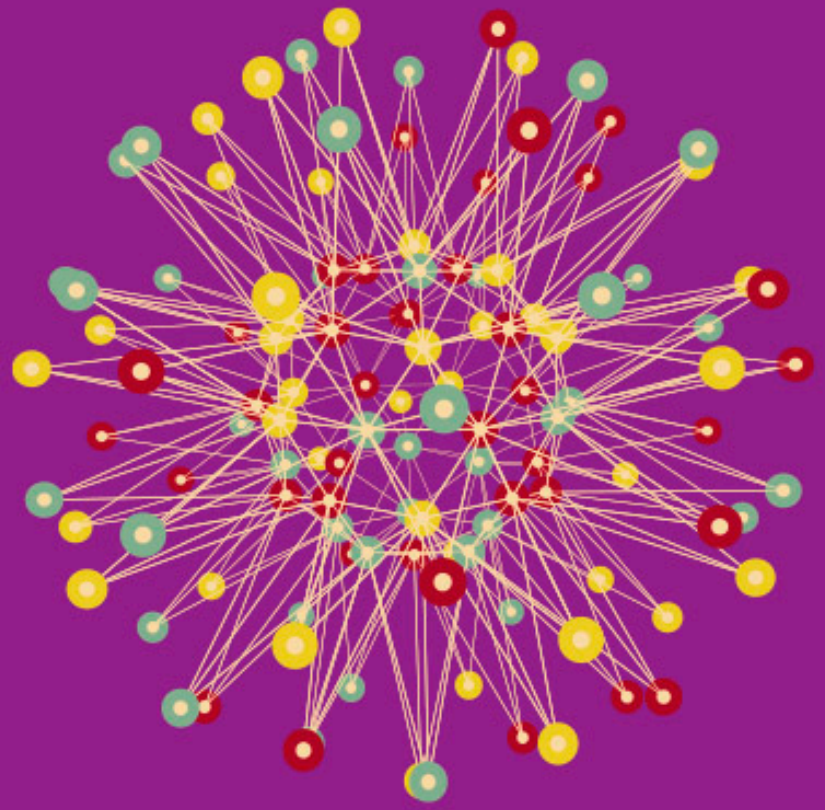
https://hal.science/hal-04448418

Submitted on 9 Feb 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution - NonCommercial - ShareAlike 4.0 International License



WEBINAIRE OALR

OPEN ACCESS LANGUEDOC-ROUSSILLON

INTRODUCTION AUX DONNÉES DE LA RECHERCHE

- DÉFINITION, ENJEUX ET CADRE JURIDIQUE
- GESTION DES DONNÉES (PLAN DE GESTION DE DONNÉES, CYCLE DE VIE DES DONNÉES,...)
- PUBLICATION DES DONNÉES (ENTREPÔTS DE DONNÉES, DATA PAPERS,...)
- UTILISATION DES DONNÉES (RECHERCHE, CITATION,...)

10 DÉCEMBRE 2019 14H00

INSCRIPTION : www.open-access.fr

COMUE LANGUEDOC-ROUSSILLON UNIVERSITÉS

Occitanie

Hensens Hanka
Responsable IST
IRD Montpellier
Membre du groupe OALR
hanka.hensens@ird.fr

Filatre Annabelle
Responsable IST
AgroParisTech Montpellier
Membre du groupe OALR
annabelle.filatre@agroparistech.fr

Points Techniques Préalables

- Pensez à :
  - Couper vos micros
  - Ouvrir un minimum d'applications (qualité de connexion)
  - Choisir, si possible, une connexion filaire

- Signaler votre présence par un bref message dans le chat (Nom et prénom)

- Poser vos questions via le chat :
  - Le conférencier y répondra lors des plages de questions prévues durant le webinaire
  - Les réponses aux questions non traitées en direct seront accessibles sur le site OALR

cc BY NC ND

Introduction aux données de la recherche : Plan

1. Contours et contexte
    1. Définition et typologie
    2. Enjeux et politiques
    3. Cadre juridique

2. En pratique
    1. Rechercher et utiliser des données
    2. Gérer des données
    3. Publier des données

Conclusion

Définitions des Données de la recherche

- Pour l'OCDE :

« **Enregistrements factuels** (chiffres, textes, images et sons), qui sont **utilisés comme sources principales pour la recherche scientifique** et sont généralement reconnus par la communauté scientifique comme **nécessaires pour valider les résultats** de la recherche. » (*Principes et lignes directrices de l'OCDE pour l'accès aux données de la recherche financée sur fonds publics* – 2006, reprise par le *Plan National pour Science Ouverte* - 2018)

- Pour la Royal Society de Londres :

« Informations **qualitatives ou quantitatives** […] qui sont factuelles. Ces données peuvent être **brutes** ou **primaires** (directement issue d'une mesure), **ou dérivée de données primaires**, mais ne sont pas encore le produit d'analyse ou d'interprétation autres que le calculs. »

Typologie des données de la recherche

- **Définition de la Commission européenne pour Horizon 2020**

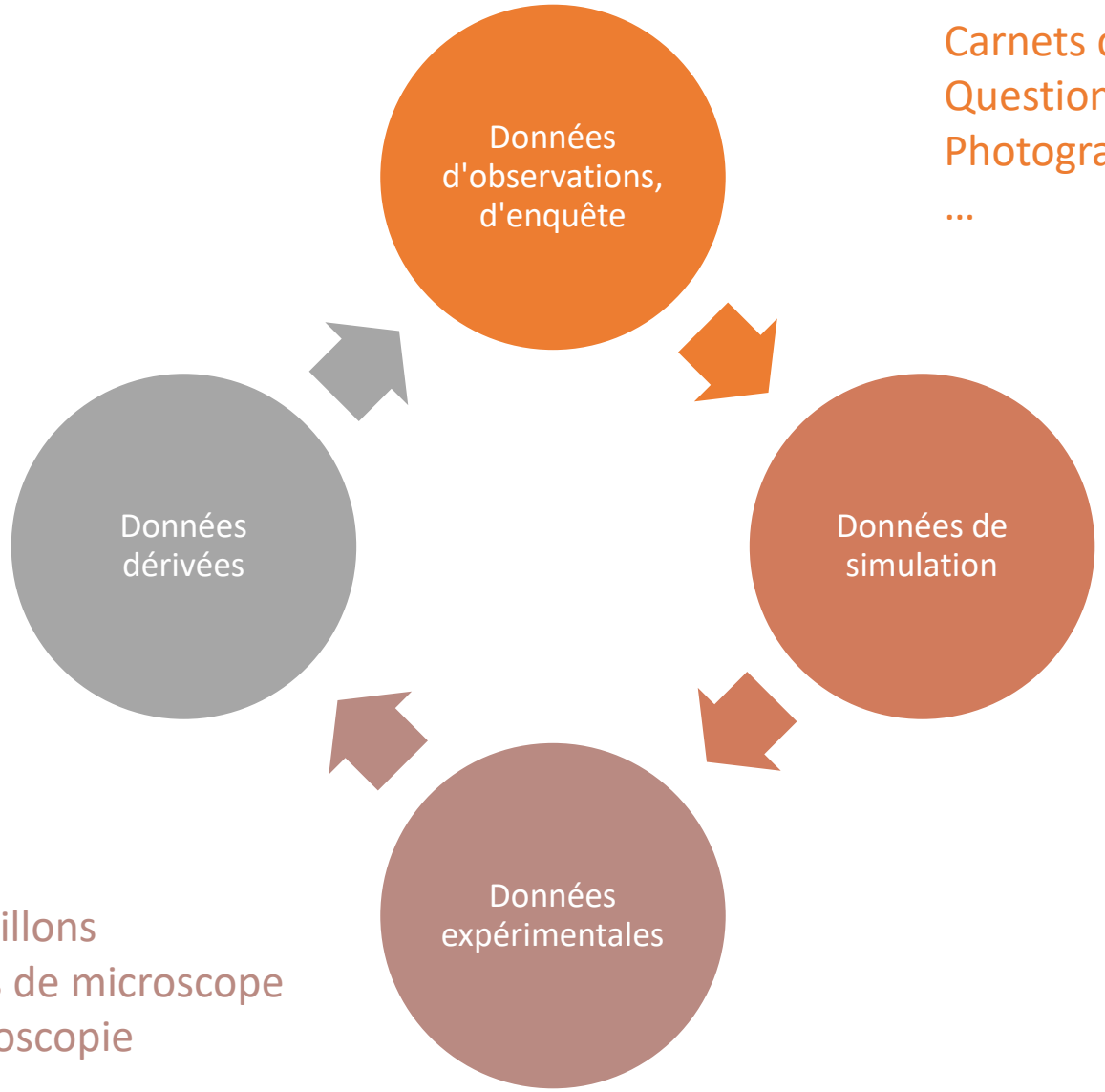
"1. Les données sous-jacente (les **données nécessaires pour valider les résultats présentés dans les publications scientifiques**), **incluant les métadonnées associées** (c'est-à-dire les métadonnées décrivant les données de recherche déposées)

2.Toute autre donnée (par exemple les données conservées qui ne sont pas directement attribuables à une publication, ou les données brutes), y compris les métadonnées associées"

Commission Européenne

HORIZON 2020

Typologie par méthode d'obtention



Carnets de terrain
Questionnaires, enquêtes
Photographies, films, fichiers audio
…

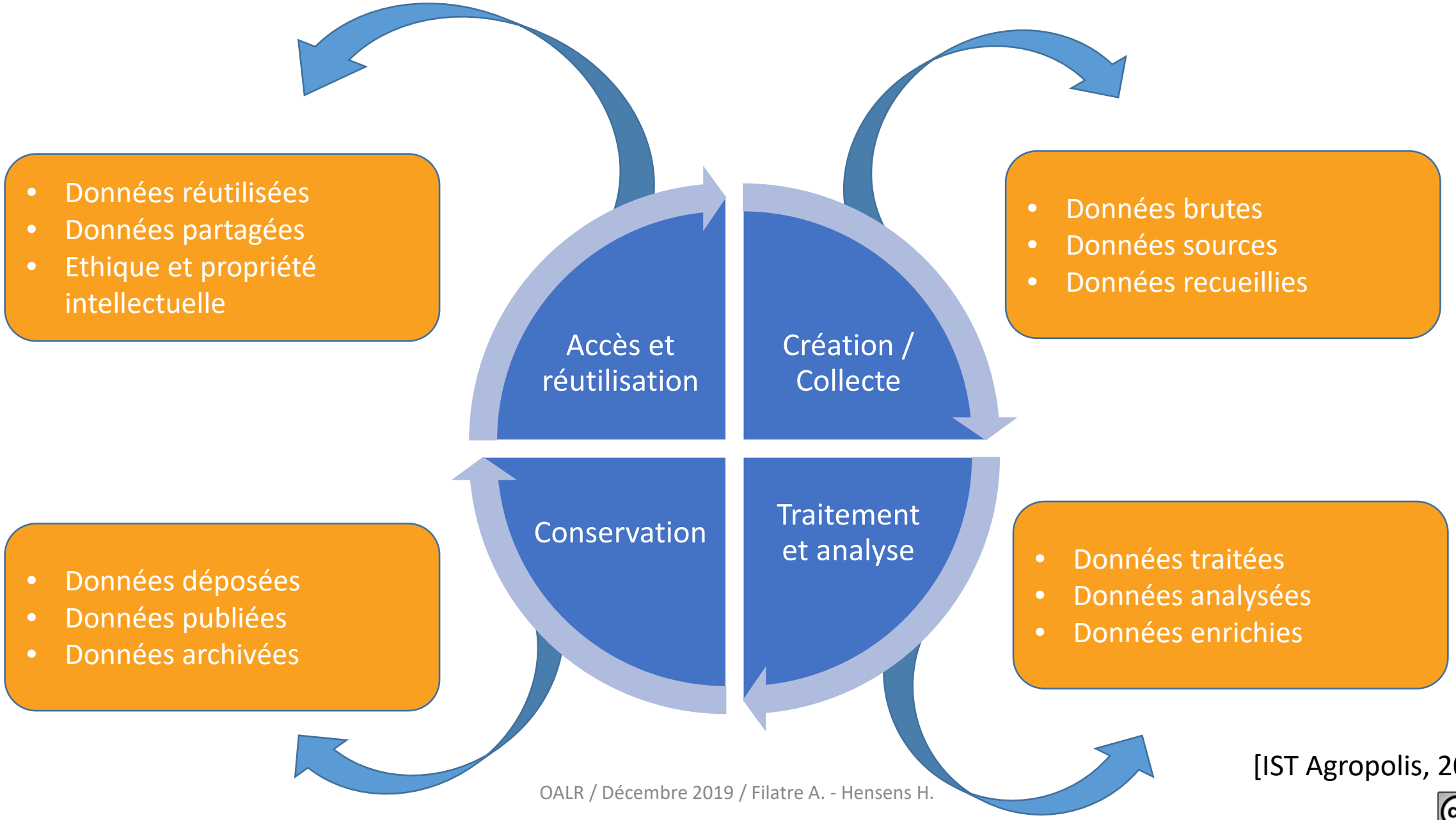
Modèles
Algorithmes
Scripts
…

Echantillons
Photos de microscope
Spectroscopie
…

Fichiers de données
Documents texte,
Feuilles de calculs
Bases de données
…

CC BY NC ND

Cycle de vie de la donnée



[IST Agropolis, 2016]

cc BY NC ND

Ne relevant pas des données de la recherche

Selon l'OCDE :

Les documents non achevés :

- les carnets de laboratoire
- les analyses préliminaires, les projets de documents scientifiques et les programmes de travaux futurs
- les examens par les pairs
- les communications personnelles avec des collègues, …

Les objets matériels :

- les souches bactériennes
- les animaux de laboratoire,…

Les productions scientifiques

- les publications scientifiques, communications à congrès,
- les supports de formation, …

Les données administratives non intégrées dans un corpus de recherche...

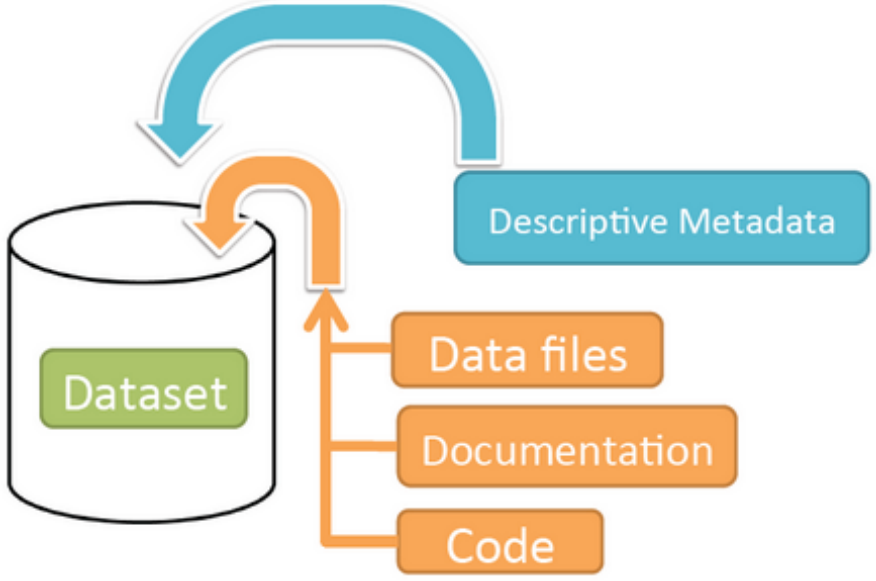
Image de Pixabay

CC BY NC ND

Définition des jeux de données (Dataset)

- « Agrégation, sous une forme lisible, de données brutes ou dérivées présentant une certaine unité, rassemblées pour former un ensemble cohérent. » (Gaillard, 2014).
- Unité de dépôt dans un entrepôt : sa granularité ne doit être ni trop petite, ni trop grande pour faciliter la citation
- Complété par des métadonnées nécessaires à leur réutilisation et à leur validation : méthodes d'obtention, logiciels nécessaires, auteurs, couverture temporelle et spatiale, etc…

Schematic Diagram of a **Dataset** in Dataverse 4.0



Container for your data, documentation, and code.

CC BY NC ND

Les enjeux de l'ouverture des données de la recherche scientifique

Enjeux scientifiques

- Défis scientifiques complexes
- Accroître l'impact de la science
- Faire évoluer les systèmes d'édition
- Eviter la perte de données

Enjeux économiques

- Valeur économique de la donnée
- Conditions de financement des bailleurs
- Continuité de la recherche
- Big Data et data science

Enjeux sociétaux

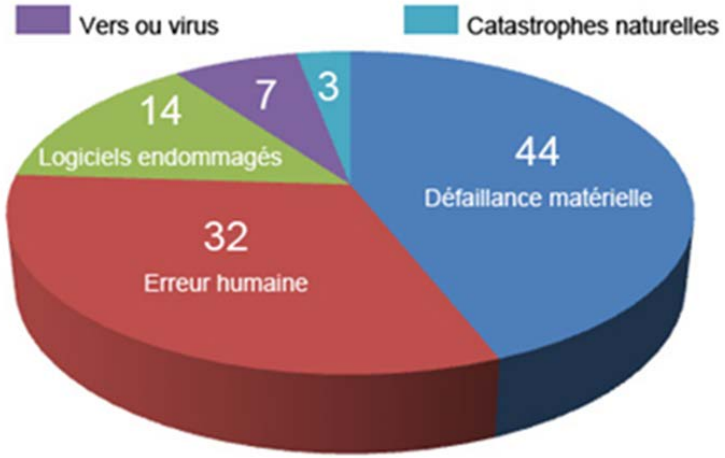
- Sécurisation des données
- Interopérabilité des données
- Participation citoyenne
- Transparence du processus scientifique

cc BY NC ND

Contexte scientifique : la perte des données

20 ans après publication,
80 % des données scientifiques sont perdues*…

- Causes :
  - Destruction des supports, virus
  - Obsolescence matérielle ou logicielle
  - Lieu de stockage indéfini
  - Erreur humaine,
    départ de personnel…



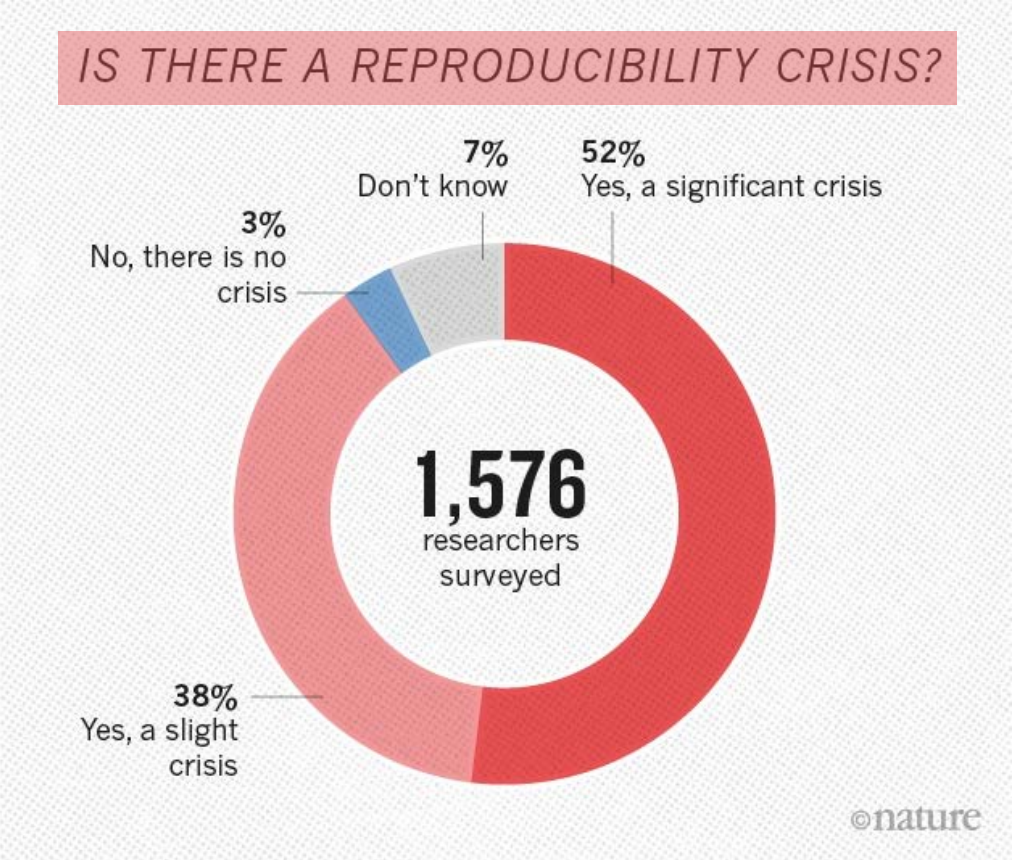
- Conséquences :
  - Perte de temps, d'argent (fonds publics…)
  - Pas de vérification des résultats possible
  - Pas de comparaison des résultats dans le temps ou l'espace
  - Pas de réutilisations par d'autres publics pour d'autres fins

** VINES Timothy H., et al. [The Availability of Research Data Declines Rapidly with Article Age](), Current Biology, 2014.*

CC BY NC ND

Contexte scientifique : la reproductibilité de la science

Les effets négatifs du « Publish or Perish »…



"More than 70% of researchers have tried and failed to reproduce another scientist's experiments, *and* ***more than half have failed to reproduce their own experiments"*** (!)

Nature, mai 2016 : https://www.nature.com/news/1-500-scientists-lift-the-lid-on-reproducibility-1.19970

cc BY NC ND

Contexte international : la Science Ouverte

Open Data

- Loi Cada (1978)
- Directive PSI (2003)
- Directive Inspire (2007)
- Charte du G8 (2013)
- Loi Valter (2015)

Open Science

- Horizon 2020 (2013-2016)
- Loi pour une République Numérique (2016)
- Plan National pour la Science Ouverte (2018)
- ANR : politique Science Ouverte (2019)

Open Access

- Initiative de Budapest (2002)
- Initiative de Berlin(2003)
- Directive Inspire (2003)
- Déclaration de l'OCDE (2006)
- Plan S (2018)

cc BY NC ND

Focus : Plan National pour la Science Ouverte

Juillet 2018

«La France s'engage pour que les résultats de la recherche scientifique soient ouverts à tous, chercheurs, entreprises et citoyens, sans entrave, sans délai, sans payement.» (https://www.ouvrirlascience.fr)

- **Axe 1 Généraliser l'accès ouvert aux publications**
- **Axe 2 : Structurer et ouvrir les données de la recherche**
- **Axe 3 : s'inscrire dans une dynamique durable, européenne et internationale**

CC BY NC ND

Focus : Politique science ouverte de l'ANR

- Favoriser le libre accès aux publications (Open Access)
- Contribuer à l'ouverture des données quand cela est possible (Open Research Data)
- Coordonner les actions aux niveaux européen et international

https://anr.fr/fr/lanr-et-la-recherche/engagements-et-valeurs/la-science-ouverte/

cc BY NC ND

Le cadre juridique français

- D’une manière générale, **les données sont réputées « de libre parcours »** : elles ne sont pas protégées par le droit d’auteur
- Les données issues de la recherche sont **considérées comme des documents administratifs,** si elles ont été produites dans le cadre d'une mission de service public ou majoritairement grâce à des fonds publics :
  - Elles sont donc **communicables à la demande** (sauf exceptions légales) si elles sont "**achevées**".
  - Depuis la Loi pour une République Numérique (2016), elles sont même soumises à un principe d’**ouverture par défaut.**
  - Ces données publiques doivent être diffusées **gratuitement et librement réutilisables**

cc BY NC ND

Les exceptions : publication sous conditions

- Les données **personnelles** ou les données **sensibles**
- Les données protégées par le **droit d'auteur** (œuvres originales)
- Les données qui impliquent **un partenaire étranger ou privé** (Droit sui generis des bases de données)
- Les données concernant les **ressources génétiques et « connaissances traditionnelles » associées** : *Protocole de Nagoya*
- Les informations pouvant avoir un impact sur la **conservation de la biodiversité** *Code de l'Environnement*
- Les données présentant des risques pour la protection du potentiel scientifique et technique de la nation, produites dans des laboratoires classés **« zones à régime restrictif » (ZRR)**

cc BY NC ND

Les exceptions : Publication interdite

- Données relatives à la **sécurité publique, sureté de l'Etat et sécurité des l'établissements** : biens, personnes, informatique, …
- **Secret professionnel** : secret des procédés, secret médical, secret de l'instruction, secret bancaire, …
- **Secret défense**

CC BY NC ND

Les licences précisent les droits de diffusion

Décret n° 2017-638 du 27 avril 2017 relatif aux licences de réutilisation à titre gratuit des informations publiques et aux modalités de leur homologation

- l'ODbL (Open Database License version 1.0), pour contrôler les redistributions et les travaux dérivés, ou pour une diffusion internationale (~CC-BY-SA)
- la Licence Ouverte Etalab, si le suivi du devenir des données n'est pas recherché et si les données sont essentiellement distribuées en France (~CC-BY)

ODbL stands for Open Data Commons Open Database License

LICENCE OUVERTE
OPEN LICENCE

Les licences Creative Commons, alternative internationale possible

- Pour choisir : https://creativecommons.org/choose/?lang=fr
- Pour comprendre : https://creativecommons.org/licenses/?lang=fr-FR

creative commons

BY NC ND

Des Questions ?

Introduction aux données de la recherche : Plan

1. Contours et contexte
    1. Définition et typologie
    2. Enjeux et politiques
    3. Cadre juridique

2. En pratique
    1. Rechercher et utiliser des données
    2. Gérer des données
    3. Publier des données

Rechercher des données : les entrepôts de données

- Un entrepôt est un service en ligne permettant le dépôt, la description, la conservation, la recherche et la diffusion des jeux de données
- Les entrepôts peuvent être généralistes ou thématiques, publics ou propriétaires, gratuits ou payants, réservés à un organisme ou ouvert à tous
- Leurs interfaces de recherche peuvent être plus ou moins avancées : moteurs de recherche interne, vocabulaires, filtres, facettes, …

zenodo

DRYAD

figshare

PANGAEA.

cc BY NC ND

Rechercher des données : les répertoires et annuaires

Les annuaires et répertoires classent les entrepôts par critères géographiques, thématiques, …

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

Annuaires d ’entrepôts certifiés

- CoreTrustSeal

cc BY NC ND

Rechercher des données : les moteurs de recherche

- Recherche fédérée sur plusieurs entrepôts
- Données de tous types et de tous les domaines
- Filtres puissants

Moteurs de recherche de données

- Datacite (DOI)
- DataSearch (Elsevier, v. Beta)
- Google Dataset Search (v. Beta)
- BASE (ressources « académiques »)

CC BY NC ND

Droits d'utilisation

- **Si une licence est associée au jeu de données, cela clarifie les droits** d'utilisation, même si ça peut les restreindre.
- **En l'absence de licence, un jeu de données peut être protégé** par le code de la propriété intellectuelle (œuvres originales, bases de données privées ayant nécessité des investissements substantiels) : mieux vaut alors demander l'autorisation du titulaire du droit…
- Dans tous les cas, sauf dans le cadre de la licence CCO, l'utilisateur de données est obligé de **citer sa source**, car le droit de paternité, en tant que droit moral, est inaliénable.

CC BY NC ND

Citer un jeu de données

Le format minimal d'une référence bibliographique d'un jeu de données comporte 5 éléments, selon la recommandation de DataCite :

Auteur (Année de publication) : Titre. Editeur (=organisme responsable). Identifiant

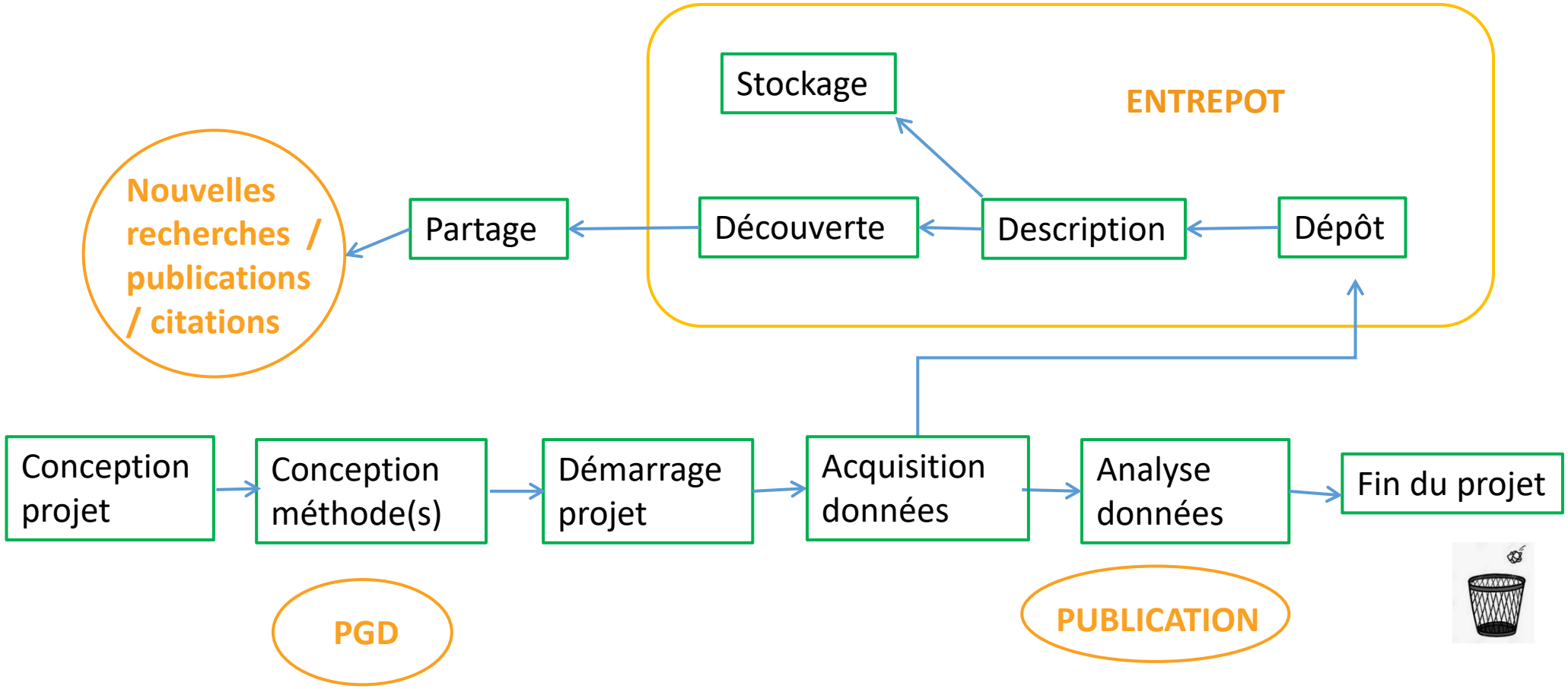
Ce format minimal peut être complété par la version et le type de ressource :

Auteur (Année de publication) : Titre. Version. Editeur. Type de ressource. Identifiant

L'outil en ligne DOI Citation Formatter développé par DataCite et CrossRef génère, à partir du DOI, la référence bibliographique du jeu de données dans plus de 500 formats au choix.

CC BY NC ND

Changer les pratiques de gestion des données



CC BY NC ND

Un PGD/DMP Qu’est-ce que c’est ?

Définition

- ✓ Un document formalisé (type formulaire)
- ✓ Liste toutes les questions à se poser
- ✓ A toutes les étapes du cycle de vie des données
- ✓ A commencer le plus tôt possible (au démarrage d'un projet)

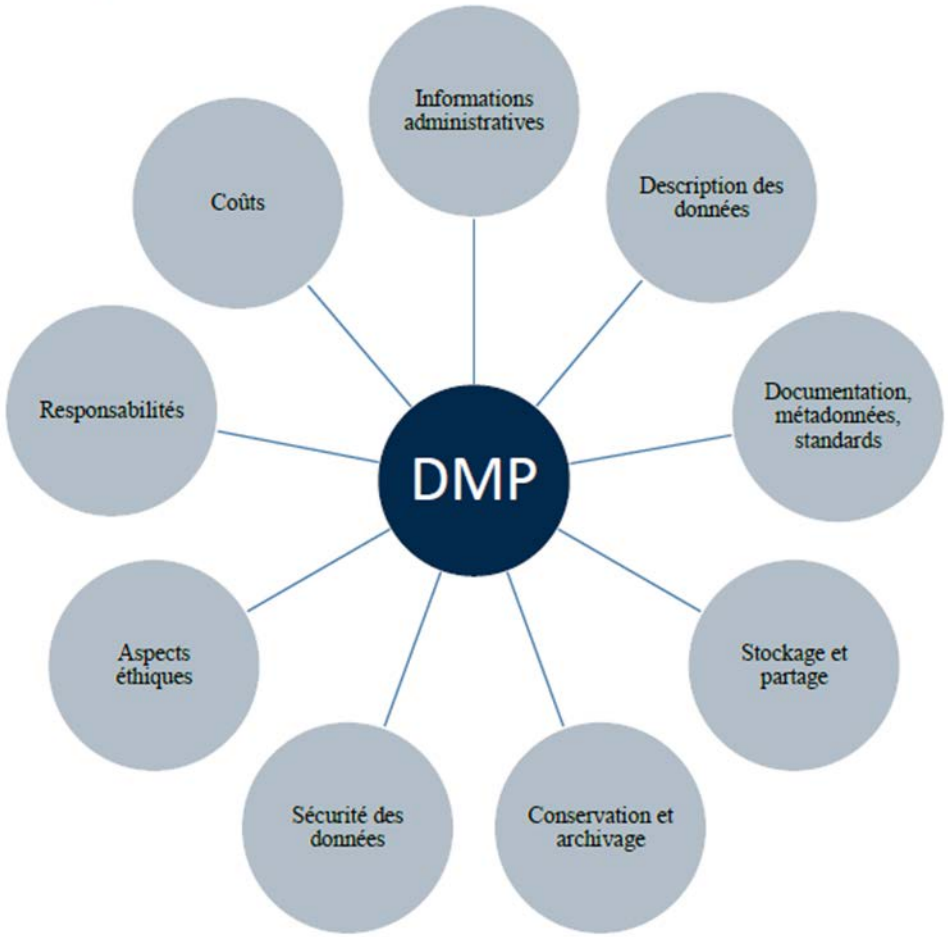
Objectifs

- ✓ Aide à la mise en place de bonnes pratiques de gestion des données
- ✓ Pour des données plus fiables, compréhensibles, disponibles tout au long du projet
- ✓ Préservées si elles le méritent pour une réutilisation future

CC BY NC ND

Les PGD ont toujours les mêmes rubriques

1. **Description du projet de recherche :** but, partenaires, responsabilités, leadership,…
2. **Inventaire et Description des données** : type, taille, méthode de collecte et traitement,…
3. **Documentation** : métadonnées, documents complémentaires, logiciel nécessaire,…
4. **Sauvegarde et sécurité** : matériel et méthode, périodicité et nombre de copies, versioning, nommage,.
5. **Droit et éthique :** selon le type de données (personnelles, santé, vivant), les lois,…
6. **Conservation** : quelles données à long terme, combien de temps, à quel endroit
7. **Partage** (pendant et après le projet) : accès, licences, embargo, confidentialité,…
8. **Ressources** : matérielles, financières, humaines, …



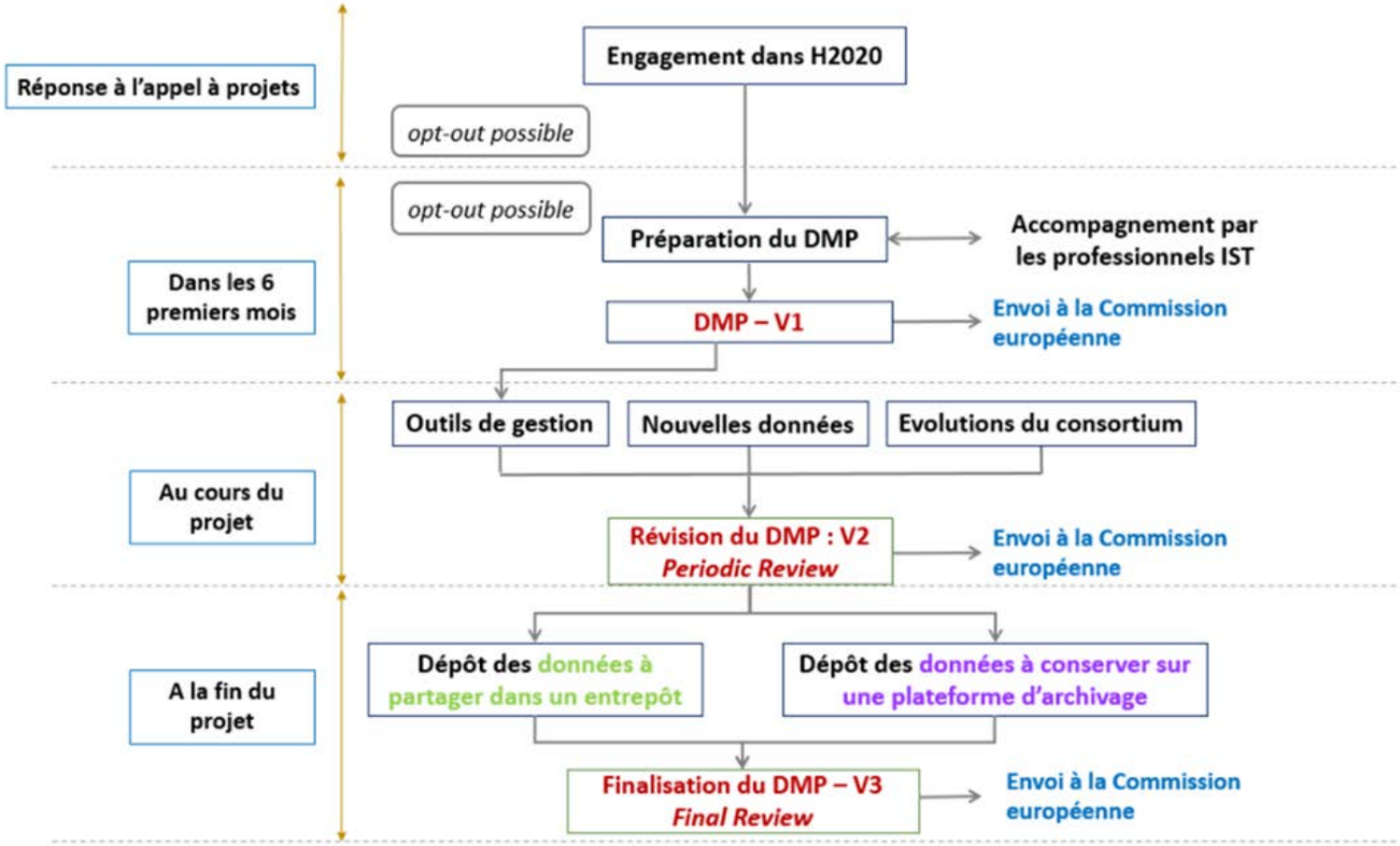
Fiche synthétique Doranum

CC BY NC ND

LE PGD est évolutif

Un PGD n'est pas figé :

- il doit être mis à jour tout au long du projet
- Il doit comporter des versions successives
- *Exemple H2020 :*
  3 versions minimum
  (pas de maximum)



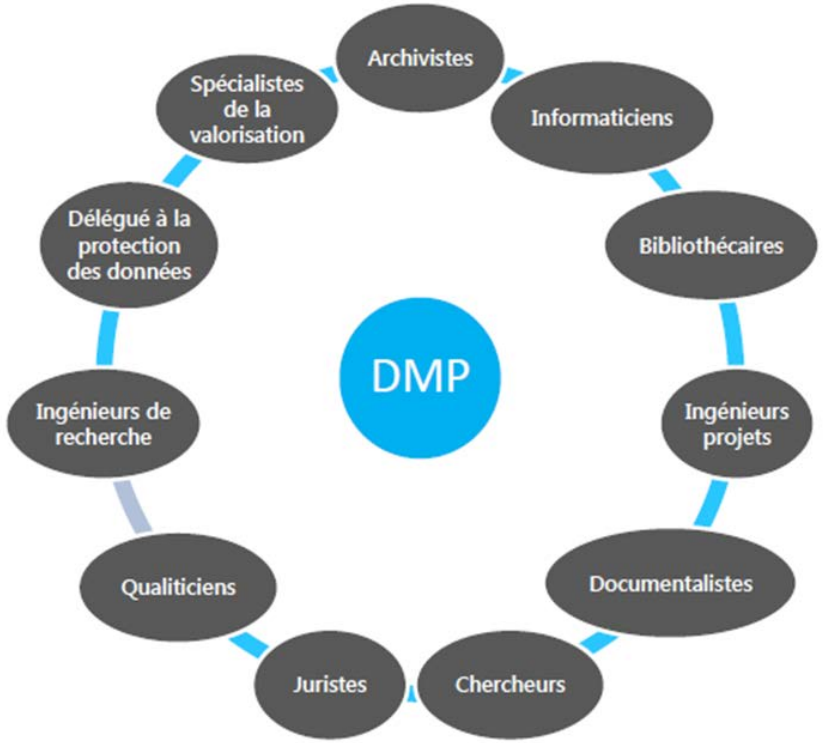
A. Cartier, M. Moysan, N. Reymonet, d'après les *Guidelines on FAIR data management*

cc BY NC ND

Le PGD est un travail de groupe

Participent à la gestion des données et donc à la rédaction du PGD :

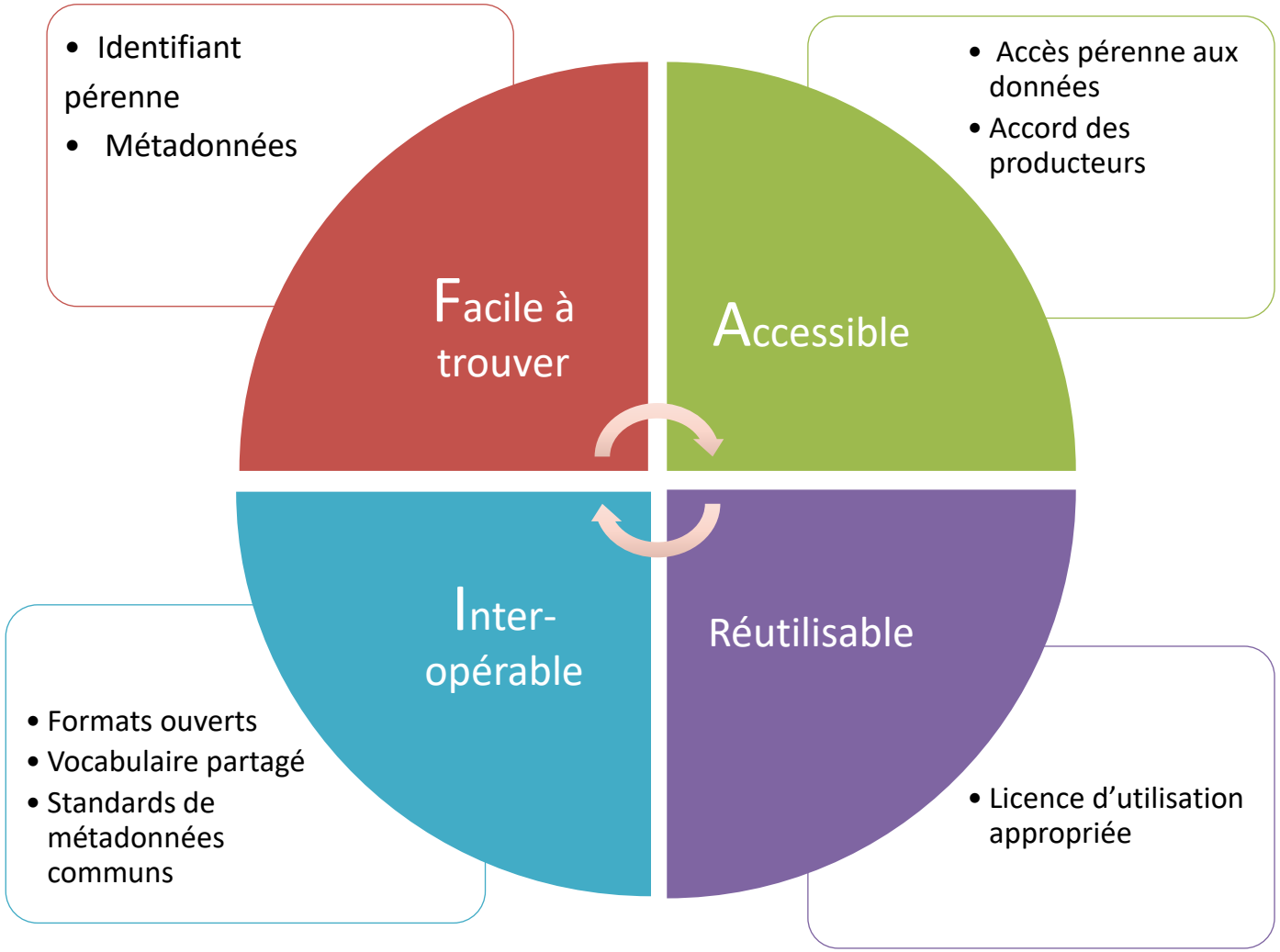
- **Scientifiques** : description, validation
- **Ingénieurs-données :** sauvegarde, qualité
- **Informaticiens :** sécurisation, infrastructure
- **Spécialistes IST :** modèles de PGD, standards de métadonnées
- **Archivistes :** données à conserver, durées et solutions techniques
- **Juristes :** propriété intellectuelle, législation, données personnelles, sensibles,…
- **Chargés de valorisation :** rédaction des conventions à signer

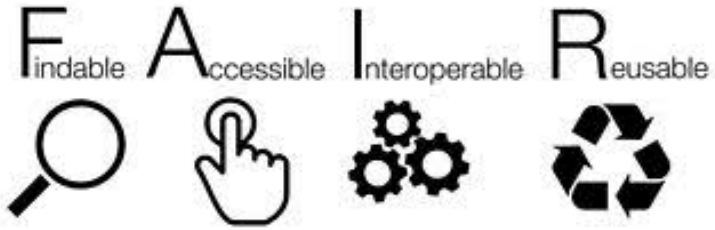


Reymonet, N., Moysan, M., Cartier, A., & Délémontez, R. (2018). Réaliser un plan de gestion de données « FAIR »

cc BY NC ND

Diffusion des données dans le respect des principes FAIR





Aussi ouvert que possible, aussi fermé que nécessaire

cc BY NC ND

Entrepôts de données : critères de choix

Caractéristiques	Fonctionnalités	Limites
Pérennité de l'entrepôt	Préservation et archivage des données garanti	Discipline(s)
Facilité de dépôt	Identifiant pérenne (DOI)	Coûts
Facilité de recherche / découverte des données	Contrôle d'accès aux données	Types de données
Localisation du serveur	Gestion des droits	Formats acceptés
Description des données _métadonnées_ (adaptation disciplinaire, utilisation de standards, contrôles,…)	Statistiques d'usage (consultation, téléchargement, traçabilité, provenance,…)	Volumes maximaux (fichier, jeu de données)
Production de citation	Versionning	Licences proposées
Organisation (arborescence,…)	Interopérabilité / moissonnage	
	Prépublication (lien privé pour reviewers,…)	Déposants autorisés (pays, organisme,…)

Principe du Data paper

Jeu de données

1	73	47	1	7.5	40
2	18	6	1	12.2	53
3	18	10	1	77.3	725
4	47	28	1	31.1	143
5	18	15	1	18.7	60
6	4	3	2	28.23	80
7	7	6	4	6.7	13
8	33	31	2	30.6	118
9	18	8	1	3.7	23
10	3	2	1	25.0	70
11	5	4	2	50.0	104
12	15	12	1	22.3	137
13	12	9	1	21.0	69
14	3	3	10	114.0	171
15	6	5	2	59.5	156
16	10	10	4	61.5	172

Décrire un jeu de données

en vue de sa réutilisation

Mettre à disposition d'une communauté scientifique

Informer la communauté scientifique

Entrepôt de données

Data paper

Data in Brief 15 (2017) 92–96

Contents lists available at ScienceDirect

Data in Brief

journal homepage: www.elsevier.com/locate/dib

Data Article

Integrated dataset of anatomical, morphological, and architectural traits for plant species in Madagascar

Amira Azizan [a], Emma Guillon [a], Yves Caraglio [c], Patrick Langbour [b], Sébastien Paradis [b], Pierre Bonnet [c], Yannick Brohard [c], Christine Heinz [c], Nabila Boutahar [b], Loïc Brancheriau [b,*]

[a] University of Montpellier, Dept. of Biodiversity, Ecology and Evolution, Montpellier, France
[b] CIRAD, UR BioWooEB, 34398, Montpellier, France
[c] CIRAD, UMR AMAP, 34398, Montpellier, France

ARTICLE INFO

Article history:
Received 7 July 2017
Accepted 5 September 2017
Available online 12 September 2017

Keywords:
Morpho-architectural traits
Plant architecture
Wood anatomy
Madagascar

ABSTRACT

In this work, we present a dataset, which provides information on the structural diversity of some endemic tropical species in Madagascar. The data were from CIRAD xylotheque (since 1937), and were also collected during various fieldworks (since 1964). The field notes and photographs were provided by French botanists; particularly by Francis Hallé. The dataset covers 250 plant species with anatomical, morphological, and architectural traits indexed from digitized wood slides and fieldwork documents. The digitized wood slides were constituted by the transverse, tangential, and radial sections with three optical magnifications. The main specific anatomical traits can be found within the digitized area. Information on morphological and architectural traits were indexed from digitized field drawings including notes and photographs. The data are hosted in the website ArchiWood (http://archiwood.cirad.fr).

© 2017 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

* Corresponding author.
E-mail address: loic.brancheriau@cirad.fr (L. Brancheriau).

http://dx.doi.org/10.1016/j.dib.2017.09.004
2352-3409/© 2017 The Authors. Published by Elsevier Inc. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

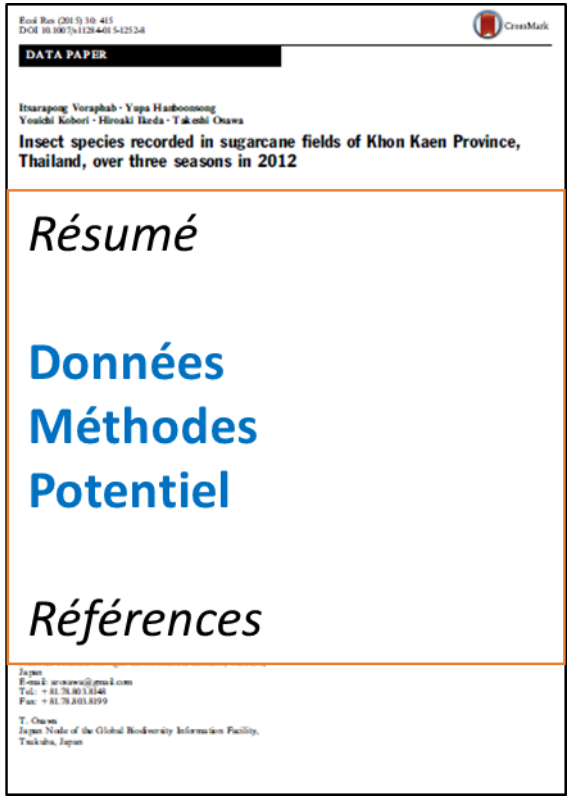
CC BY NC ND

Contenu et Objectif du Data Paper

Un article scientifique… pas comme les autres !

- Décrit un jeu de données
  - → pour permettre leur réutilisation et leur identification
- Détaille les méthodes d'obtention
  - → pour permettre de reproduire l'étude
  - → pour montrer rigueur scientifique/qualité données
- Montre le potentiel de réutilisation des données
  - → originalité des données
  - → pour qui et pour quoi elles peuvent être utiles
- Donne accès aux données

Rubriques types :


Eest Res (2015) 30: 415
DOI 10.1007/s11284-015-1252-4

DATA PAPER

Itsarapong Voraphab · Yupa Hanboonsong
Youichi Kobori · Hiroaki Ikeda · Takashi Osawa

Insect species recorded in sugarcane fields of Khon Kaen Province, Thailand, over three seasons in 2012

Résumé

Données
Méthodes
Potentiel

Références

Japan
E-mail: [illegible]
Tel.: +81.78.803.8148
Fax: +81.78.803.8199

T. Osawa
Japan Node of the Global Biodiversity Information Facility,
Tsukuba, Japan

Ni résultats, ni analyses, ni discussion

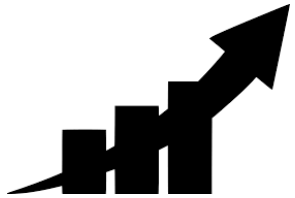
Partage de données : entre injonctions …

Incitation ou obligation de partage, de PGD,…
des bailleurs, des politiques, des éditeurs, des établissements,…

CC BY NC ND

… et bénéfices

Augmenter sa visibilité et ses citations
Le partage des données augmente l'attractivité de l'article



Étendre son réseau
nouvelles opportunités pour échanger/collaborer

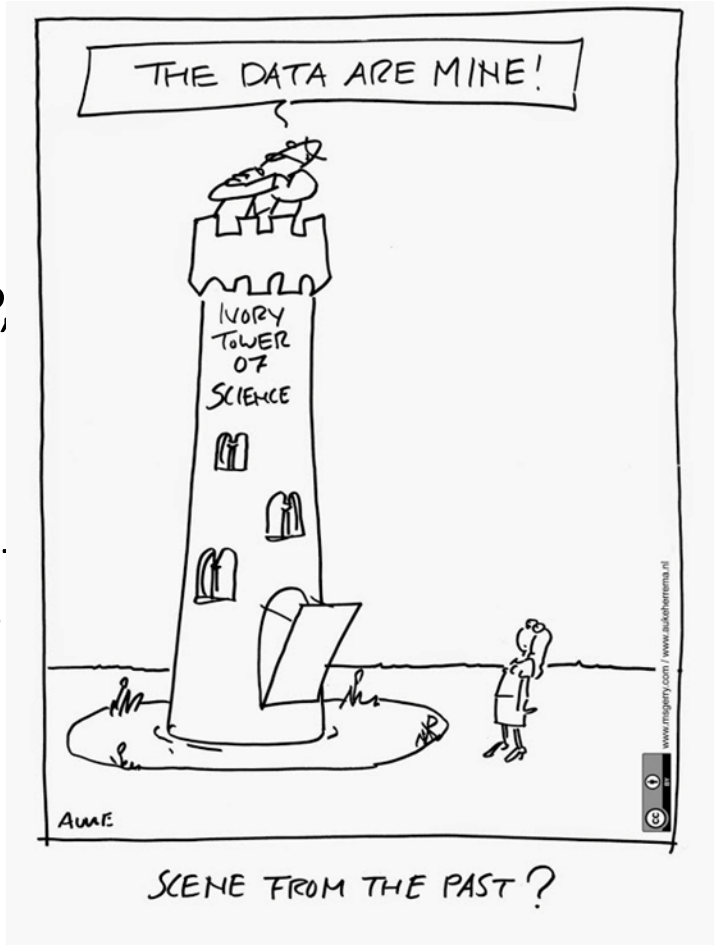
Améliorer sa réputation
et attirer sur son profil chercheur

Renforcer ses possibilités de recevoir des financements
(compatibilité avec les exigences des financeurs)

CC BY NC ND

Lever les freins à l'ouverture des données

- **C'est à moi ?…**
  - Les données de la recherche ne relèvent pas du champs de la propriété intellectuelle, …
- **C’est compliqué ?…**
  - Utiliser les standards de votre discipline, un modèle de PGD/DMP, les sites de référence, les services d’appui, travailler en équipe,…
- **C’est cher ?…**
  - Les frais (de stockage, d'organisation, de dépôt dans un entrepôt - la plupart sont gratuits - ...) sont éligibles dans les appels à projets
- **C’est risqué ?…**
  - Des sites vous aident à repérer les exceptions à la règle générale d'ouverture des données… et les services d’appui peuvent être contactés au moindre doute !



CC BY NC ND

Sources

- Ateliers du Colloque Science Ouverte au Sud, Dakar (Sénégal), 23-25/11/2019 : https://opensciencesud.sciencesconf.org/resource/page/id/6
- Ciolek-Figiel A. et Cosserat F., 2016. Gestion et diffusion des données de la recherche, Vandoeuvre-lès-Nancy : Inist – CNRS. Disponible à : https://drive.google.com/file/d/0BxKZLWq08xX-TW5VOEUtd2FSRE0/view
- CoSO https://www.ouvrirlascience.fr/college-donnees-de-la-recherche/ Livrables & productions (en particulier, cadre juridique – 2017)
- CoopIST CIRAD : https://coop-ist.cirad.fr/gerer-des-donnees
- Datapartage INRA : http://www6.inra.fr/datapartage/
- Fabre C. 2019. Gestion et diffusion des données de la recherche – Urfist Toulouse – 17 & 18 juin 2019, 82 p.
- Fiches pratiques CIST Agropolis : http://www.ist.agropolis.fr/les-ressources/fiches-pratiques
- IRD Data : https://data.ird.fr/
- Lafosse, 2018. Initiation aux données de la recherche – URFIST Paris – 27 & 28 mars 2019
- Maurel L. 2018. La réutilisation des données de la recherche après la loi pour une République numérique. *La diffusion numérique des données en SHS - Guide de bonnes pratiques éthiques et juridiques*, Presses Universitaires de Provence, 9791032001790. ⟨hal-01908766⟩
- Séminaire de lancement de DataSuds, Montpellier (en particulier intervention de Lionel Maurel sur la cadre juridique des données de la recherche) : https://data.ird.fr/seminaire-de-lancement-de-lentrepot-datasuds/
- Serres : https://guides-formadoct.ubretagneloire.fr/donnees_recherche
- Slide Share IST IRD : https://fr.slideshare.net/IST_IRD/presentations
- Sites d'aide à la décision pour le partage des données : : questionnaire INRA, logigramme de l'ENPC

cc BY NC ND

Des Questions ?

Merci à toutes et à tous