

Plan de gestion des données

Comment ?

Caroline DOUCOURE, Hanka Hensens
(IST-IRD)

dmp@ird.fr

Supports : <https://tinyurl.com/2p8ts5pp>



IRD-IST, Plans de gestion des données, 2022





Plan de gestion des données

Comment ?

DMP et cycle de vie des données

Le modèle Science Europe/ANR en détail

Recommandations, aides et assistance

Le DMP accompagne le cycle de vie de la donnée

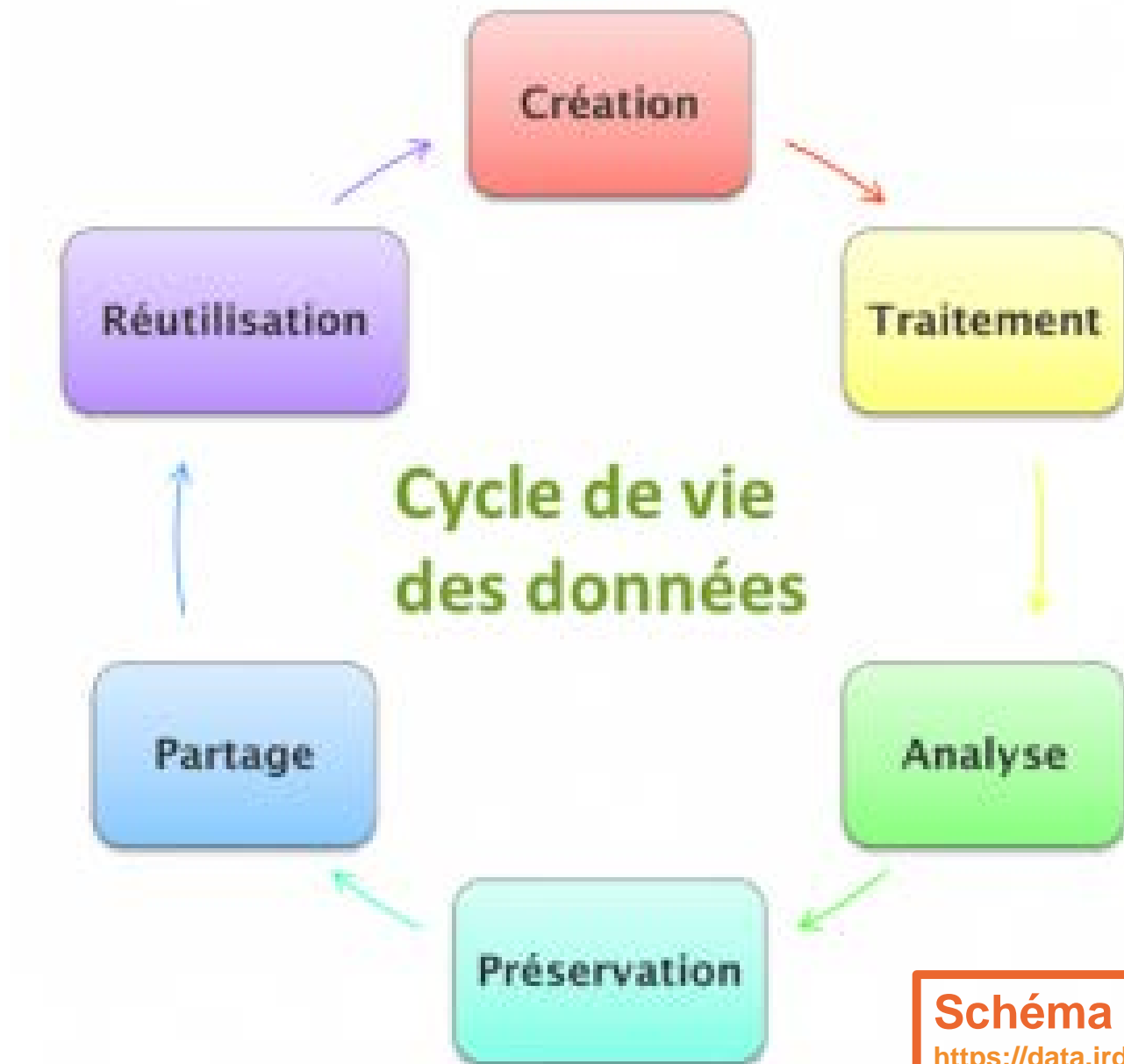


Schéma du cycle de vie des données

<https://data.ird.fr/utiliser-des-donnees/>

Le DMP permet de se poser les bonnes questions

- Quels sont les besoins générés par la gestion des données ?
- Comment seront-ils financés, à court et à long terme ?

Ressources

- En quoi consiste le projet ?
- Qui sont les partenaires ?
- Quelle est la politique de gestion des données ?
- Qui est responsable de la gestion des données ?

Responsabilités dans le projet

- Quelles données seront produites/utilisées au cours du projet ? (type, format, volume et accroissement...).
- Comment seront-elles produites ou transformées ?

Collecte des données

- Comment les données seront-elles identifiées, décrites ?
- Quels standards de métadonnées utilisera-t-on ?

Documentation des données

- Qui pourra accéder aux données ?
- Quelles données seront partagées ?
- Avec qui ? Comment ?
- Dans quel délai ?
- Sous quelle licence ?

Accès et partage des données



- Comment, où, par qui, seront stockées, sauvegardées et sécurisées les données au cours du projet ?

Sauvegarde des données

- Quel plan pour l'archivage et la préservation à long terme ?
- Pour quelles parties des données ?

Archivage et préservation des données

- Qui sera propriétaire des données produites ?
- Quels sont les droits sur les données externes réutilisées ?

Propriété intellectuelle

- Des données sensibles seront-elles produites ou utilisées ?
- Comment sera assurée leur protection/anonymisation ?

Ethique



Plan de gestion des données

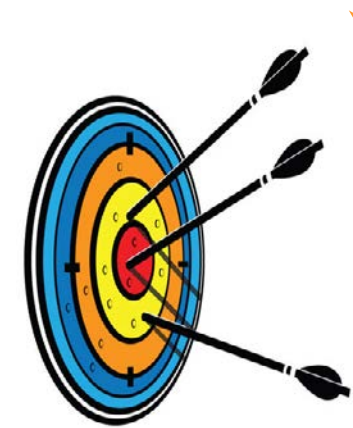
✓ DMP et cycle de vie des données

❑ Le modèle Science Europe / ANR

1. Description des données collectées ou réutilisées
2. Documentation et qualité des données
3. Stockage et sauvegarde pendant le projet
4. Exigences légales et éthiques, codes de conduite
5. Partage des données et conservation à long terme
6. Responsabilités et ressources

❑ Recommandations, aides et assistance

6 parties, pourquoi ?



 publicdomainvectors.org

- 1. Décrire les données → **Assurer la reproductibilité des expériences**
- 2. Clarifier le cadre juridique et éthique → **Respecter les droits et personnes**
- 3. Garantir la compréhension → **Permettre la réutilisation des données**
- 4. Assurer le stockage → **Eviter les pertes**
- 5. Spécifier les modalités de partage → **Clarifier les droits de réutilisation**
- 6. Définir les responsabilités → **Assurer la bonne gestion des données**

1-Description détaillée des produits de recherche...



Présenter l'ensemble des données réutilisées ou produites
en précisant de manière synthétique (de préférence dans un tableau) :

- le type de données,
- le mode d'obtention,
- Le(s) format(s) des fichiers (de travail/de partage),
- le volume (un ordre de grandeur)

...

Pour vous aider :

<https://data.ird.fr/gerer/quelles-donnees>

Formats de fichiers ouverts/fermés (Dorandum) : de préférence non-propriétaires, ouverts, pérennes et gratuits (réutilisation des données)

Types de fichiers (Datapartage INRAE)



... dans un tableau de préférence

○ Données réutilisées

Nom abrégé ou N°	Type de données	Format de fichier	Volume estimé	Origine (producteur, DOI, accord...)	Description du contenu

○ Données produites

Nom abrégé ou N°	Type de données	Format de fichier	Volume estimé	Origine (mode de collecte, traitement, logiciels,...)	Description du contenu



Tout ce qui peut permettre de retrouver et de réutiliser les données (par d'autres ou... par vous au bout de quelques années)

- Standards de métadonnées généraux ou spécifiques à une discipline scientifique et des outils d'aide à la production de métadonnées : <https://data.ird.fr/documenter-ses-donnees/>
- Documentations associées : data dictionary, code dictionary, formulaires vierges...
- Méthodologie de collecte et traitement



Expliquer comment les données seront contrôlées et documentées durant tout le projet : <https://data.ird.fr/qualite-des-donnees/>

1. Présenter les objectifs à atteindre
2. Citer les processus qualité utilisés
3. Définir nommage et organisation (Datapartage INRAE)
4. Préciser les outils pratiques utilisés (checklists, cahiers de labo* ...)
5. Décrire le suivi régulier (hebdomadaire, mensuel ...) mis en place

* Cf guide des bonnes pratiques de Doranum pour tenir un cahier de labo : <https://www.dataacc.org/bonnes-pratiques/utiliser-un-cahier-de-laboratoire-numerique/ressource-doranum-cahiers-de-labo>



2. DOCUMENTATION ET QUALITÉ DES DONNÉES



Par exemple

- Les données issues de l'imagerie médicales seront décrites selon la norme et le format d'interopérabilité DICOM.
- Il n'existe pas à notre connaissance de standards de métadonnées pour décrire les données produites mais un modèle sera mis au point et documenté dans le cadre du projet. Les outils et logiciels de gestion développés dans ce contexte seront également documentés et autant que possibles rendus accessibles.
- Chaque retranscription d'entretien sera accompagnée d'un rapport d'entretien incluant : le contexte de passation de l'entretien, les caractéristiques sociodémographiques de la ou des personnes interrogées, un résumé et des mots-clés.
- Chaque cahier de terrain scanné est classé dans un dossier numéroté préfixé par les initiales de l'auteur type < XY_Cahier_n >.
- Chaque page numérisée est nommée d'après le schéma suivant : <initiales de l'auteur>_<année de production du cahier>_< ?? >_page_<n° de page> / <producteur_sujet_type de document_AAAAMMJJ_version.extension >
- Les entretiens retranscrits seront chacun affectés d'un numéro en fonction de l'enquête ou de la sous-enquête à laquelle il correspond et du pays et ville de collecte.



1. Procédure de sauvegarde au cours du projet :
périodicité, automatisation, nombre de copies et de versions, localisation
cf. <https://data.ird.fr/sauvegarder-ses-donnees/>
 - « 3.2.1 » : sauvegarde au moins 3 fois sur au moins 2 supports différents, dont 1 en ligne, de 3 versions successives
 - De préférence sur sites institutionnels : pérennes, régulièrement dupliqués, sécurisés,...
2. Sécurité des données, et protection des données sensibles :
procédures d'accès aux données (mots de passe, numéros IP ...),
de transfert, cryptage,...



3. STOCKAGE ET SAUVEGARDE PENDANT LE PROCESSUS DE RECHERCHE



Par exemple

- Le projet BFF alliant à la fois des partenaires privés et publics et étudiant plusieurs plantes, il sera nécessaire de détailler l'accès aux données selon les différents cas de figure en se référant à l'accord de consortium.
- Les fichiers de données sont diffusés aux équipes de recherche productrices par une plateforme d'accès à distance sécurisée. Un comité d'accès aux données sera mis en place.
- Les données seront stockées sur un espace créé sur le Cloud Sécurisé du CNRS selon le dispositif Espace MyCoRe.
- Concernant l'étude du Sorgho au Cirad, les données GBS seront accessibles aux scientifiques du partenaire Cirad.
- Les données seront traitées et gérées dans un environnement sécurisé hors réseau en utilisant la technologie du bureau virtuel.

4a-Cadre juridique français et européen



- **Rédaction de DMP** : Décret n° 2021-1572 du 3 décembre 2021 relatif au respect des exigences de l'intégrité scientifique (article 6)
- **Les données sont réputées « de libre parcours »** : elles ne sont pas protégées par le droit d'auteur
- **Les données issues de la recherche**, si elles ont été produites dans le cadre d'une mission de service public ou majoritairement grâce à des fonds publics, et si elles sont « achevées » :
 - étaient **communicables à la demande**
 - sont désormais **ouvertes par défaut**
 - doivent être **diffusées gratuitement et librement réutilisables**

Cf. Loi pour une République Numérique (2016)

Sauf exceptions légales...

4a-Exceptions - Publication sous condition



1. Les données **personnelles** ([RGPD](#)), en particulier **sensibles** ([CNIL](#)), dont les **données de santé** (secret médical) – cf [IRD Data](#)
2. Les données concernant les **ressources génétiques et « connaissances traditionnelles » associées** (partage équitable des bénéfices, contreparties) : [Protocole de Nagoya](#)
3. Les informations pouvant avoir un impact sur la **conservation de la biodiversité** [Code de l'Environnement](#)
4. Les données qui impliquent **un partenaire étranger ou privé** ([Droit sui generis des bases de données](#))
5. Les données protégées par le [droit d'auteur](#) (œuvres originales : photos, écrits...)

➤ <https://data.ird.fr/cadre-juridique/>



Droit d'auteur :

- Loi N° 2005-30 du 05 avril 2006 relative à la protection du droit d'auteur et des droits voisins

Données publiques : pas d'obligation d'ouverture

- Loi 2015-07 du 22 janvier 2015 portant Code de l'information et de la communication
- Loi n° 2017-20 du 20 avril 2018 portant Code du numérique

Données personnelles :

- Loi n° 2009-09 portant protection des données à caractère personnel
- Autorité de Protection des Données à Caractère Personnel

➤ **Législation par pays fournisseur :** <https://absch.cbd.int/fr/countries>



Si vos données peuvent soulever des problèmes éthiques ou déontologiques :

- outil d'aide ([Data Ethics Canvas de l'Open Data Institute](#)),
- préciser quel(s) comité(s) sera contacté(s)
 - Au Bénin : [Comité National d'Ethique pour la Recherche en Santé \(CNEERS\)](#)
 - Des organismes de recherche
 - Des bailleurs ([Fonds National pour l'Environnement et le Climat](#))
 - Des pays partenaires...



4. EXIGENCES LÉGALES ET ÉTHIQUES, CODES DE CONDUITE



Par exemple

- Les enregistrements audio des entretiens seront conservés le temps de leur transcription et de leur pseudonymisation puis seront détruits au plus tard à la fin du projet. Une procédure d'anonymisation des retranscriptions pseudonymisées des entretiens sera mise en place : les rapports d'entretiens seront repris un à un et tous les passages contenant des informations susceptibles d'identifier la personne interrogée ou celles qu'elle mentionne seront supprimés (avec indication de la suppression). En cas de doute, la suppression sera favorisée. Chacun de ces rapports d'entretien sera validé par l'un des chercheurs ayant conduit l'enquête.
- Les données nominatives et les données d'enquêtes sont stockées dans deux systèmes d'information différents. Le traitement pour les données nominatives du panel a été enregistré au registre CIL de l'INED sous le numéro 2012-CIL-0012
- Les droits de propriété intellectuelle sur les données du projet et leurs dérivés restent acquis par les différentes parties ayant créé ou produit ces données. Cependant, les différents membres du consortium se sont engagés à partager les données produites entre eux.

“aussi ouvert que possible,
aussi fermé que nécessaire”

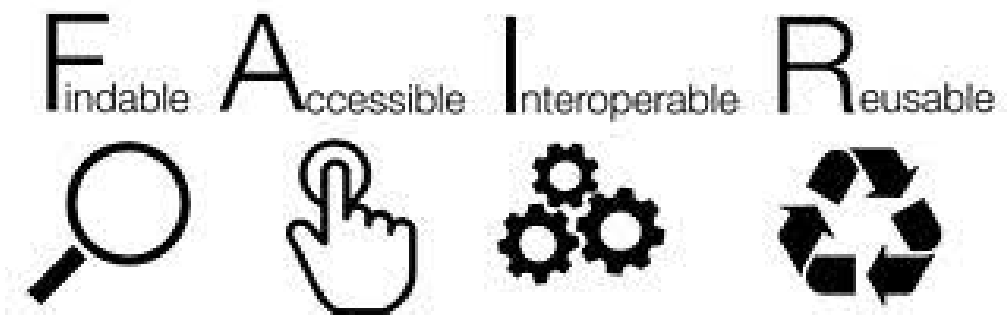
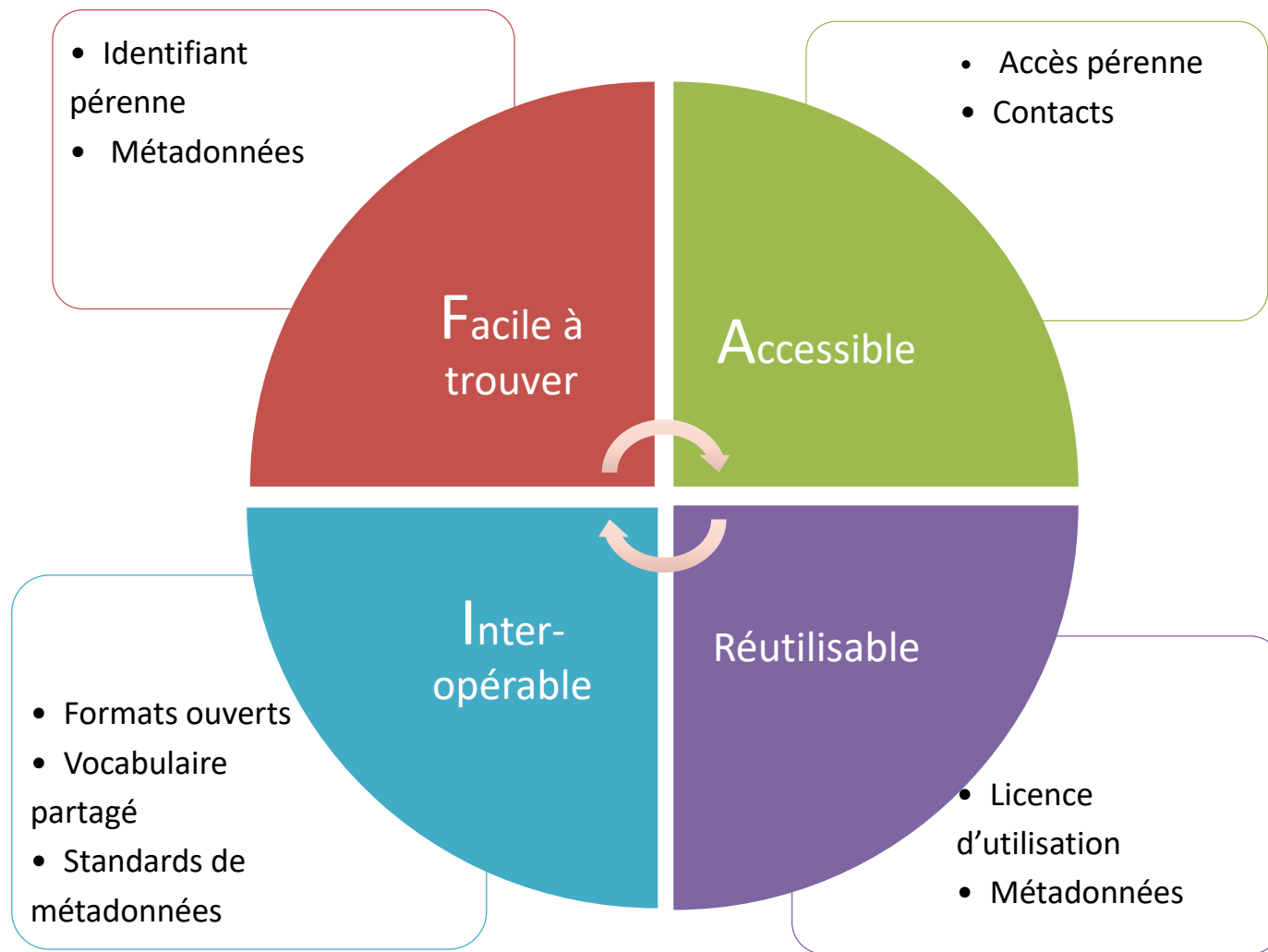


**AUSSI OUVERT
QUE POSSIBLE**



**AUSSI FERMÉ
QUE NÉCESSAIRE**

5-Diffusion et Partage : Les Grands Principes



Principes FAIR : quels enjeux et bénéfices (Doranum)

5-Partage et conservation à long terme APRES LE PROJET



1. Quelles données **conserver** ? (DCC)

2. Mes données sont-elles **partageables** ?

Aide à la décision sur IRD Data

3. **Où publier/déposer :**

- mes **logiciels** ? Software Heritage (Inria)
- mes **données** ?

Comment choisir un **entrepôt** ? + voir IRD Data

+ Rechercher des entrepôts de confiance dans l'annuaire

Re3data : <https://www.re3data.org/browse/by-subject/>

4. Choisir une **licence** (Datapartage INRAE)

Browse by subject








5-Des licences pour préciser les droits

En France, 2 choix de licences possibles pour les données publiques :
la licence ouverte Etalab et l'Open Database License (ODbL)

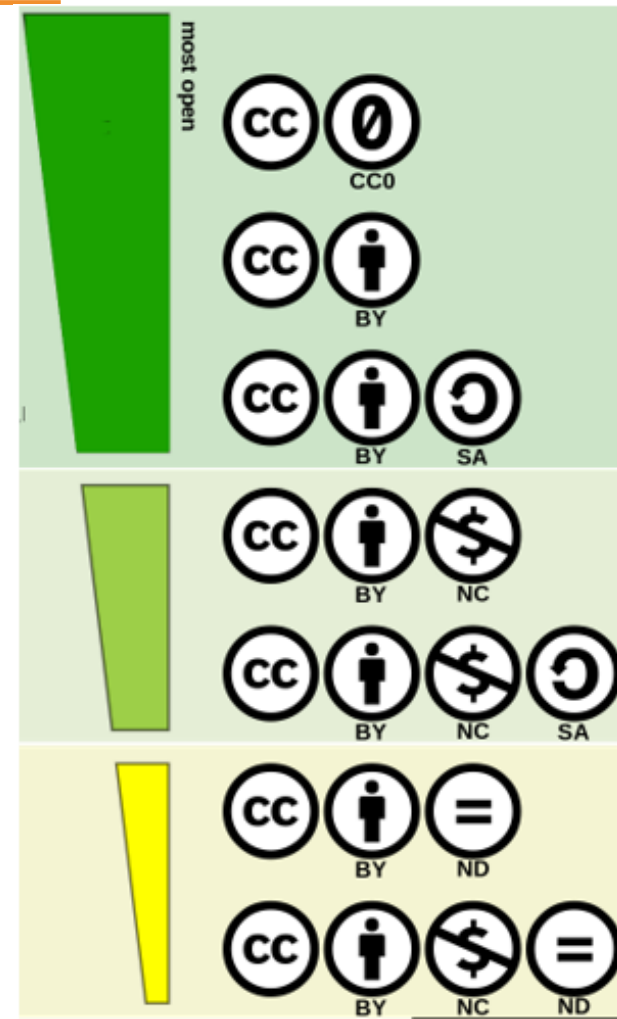
Alternative internationale : Creative Commons

Pour choisir : <https://creativecommons.org/choose/?lang=fr>

5 icônes (droits combinables) => 7 licences

	ZERO	Domaine public
	BY	Attribution = Paternité
	SA	Share Alike = Partage dans les mêmes conditions
	NC	Non Commercial
	ND	No Derivatives = Pas de Modification

Cf. https://data.ird.fr/cadre-juridique/#Choix_de_licence





5. PARTAGE DES DONNÉES ET CONSERVATION À LONG TERME



Par exemple

- Seules les données sources ainsi que les données venant en appui des résultats publiés seront conservées pour une durée minimum de 15 ans. L'ensemble des données intermédiaires seront supprimées dans l'année suivant la fin du projet.
- Les fichiers de données anonymisées seront déposés dans la plateforme DataSuds.
- Les données relatives aux sciences de la terre et environnementales ont été diffusées via l'entrepôt ouvert PANGAEA et feront également l'objet d'un "data paper" dans la revue X.
- Les données qui feront l'objet d'un archivage en libre accès seront distribuées sous licence Creative Commons - Attribution - Partage dans les Mêmes Conditions 4.0 International
<https://creativecommons.org/licenses/by-sa/4.0/>



- **Décrire les rôles et les responsabilités dans l'équipe projet** concernant les activités de gestion des données
- Si possible **nommer** la(es) personne(s) responsable(s), **prévoir des suppléants** pour les rôles critiques
- Préciser comment s'effectue la **coordination** et les modalités de décision entre partenaires
- Pour les gros projets, avec de nombreux partenaires, dans plusieurs pays, recruter **un datamanager**

6b-Ressources nécessaires à la (bonne) gestion des données



- Ne pas oublier de chiffrer : frais de stockage, coût du matériel et des logiciels, temps de personnel, formations, coûts de préparation des données pour le partage, frais d'entrepôt et d'archivage
Cf. <https://www.ukdataservice.ac.uk/manage-data/plan/costing>
- Mentionner les services obtenus gratuitement
- Estimer le salaire du/des gestionnaire(s) de données :
<https://www.openaire.eu/how-to-comply-to-h2020-mandates-rdm-costs>
- Anticiper le coût élevé de l'anonymisation des données si besoin

Coûts éligibles au financement de certains bailleurs



6. RESPONSABILITÉS ET RESSOURCES



Par exemple

- Un ETP de [n]% a été budgété pour couvrir la gestion, la documentation et la curation des données [montant].
- Des formations spécifiques en matière d'analyse et de curation de données ont été prévues en vue de la gestion du "big data" [montant].
- Coût de la maintenance de la base de données et de création d'un site web.



Plan de gestion des données

- ✓ DMP et cycle de vie des données
- ✓ Le modèle Science Europe / ANR

1. Description des données collectées ou réutilisées
2. Documentation et qualité des données
3. Stockage et sauvegarde pendant le projet
4. Exigences légales et éthiques, codes de conduite
5. Partage des données et conservation à long terme
6. Responsabilités et ressources

□ Recommandations, aides et assistance



- **Suivre les conseils de sites de référence :**
[IRD Data](#), [Coopist CIRAD](#), [Datapartage Inrae](#), [Doranum](#), [DCC](#)
- **Respecter les standards de sa communauté** ([Science Europe](#))
- **Répondre aux demandes du financeur**
(modèle, calendrier, grille d'évaluation ...)
- **Consulter des exemples de DMP :** [DMP OPIDoR](#), [DMP primés](#), [DCC](#), [DMPTool](#), [Journal Research Ideas and Outcomes \(RIO\)](#)
- **Utiliser un modèle :** https://dmp.opidor.fr/public_templates/
- **Travailler en groupe, consulter des experts**

Toutes les aides en ligne sur la [Bouée de sauvetage](#)

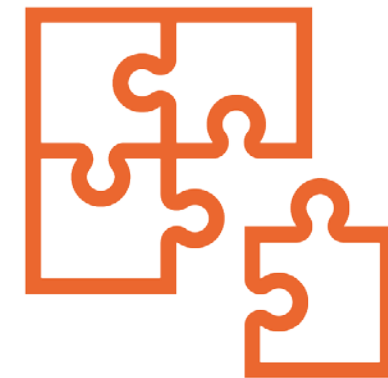


- **Choix de modèles** de financeurs et d'organismes
- **Exemples**, DMP Publics
- Accès institutionnel avec personnalisation
- **Recommandations** en ligne du modèle/d'autres organismes
- Outil **collaboratif** : rédaction en équipe, commentaires, partage
- **Téléchargements** (pas de versionning)

<https://dmp.opidor.fr>



La minute Plan de gestion de données (2'57'')



Des questions ?

