

Le plan de gestion de données

Data management plan

Laurence Leemput et Cécilia Jullien / BU, Service d'appui à la recherche

28 novembre 2024 / 12h

Science ouverte : enjeux 1/2

Trois axes :

- **Généraliser l'accès aux publications.** Les publications issues de fonds publics devront obligatoirement être diffusées en open access. Le rôle de l'archive ouverte HAL est réaffirmé.

Toutes les publications des chercheurs de CY sont à déposer dans HAL.

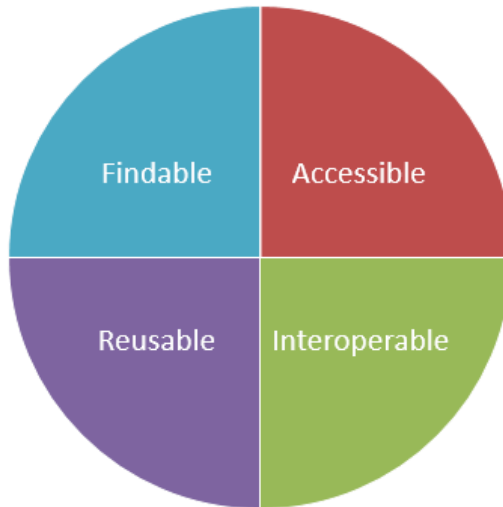
- **Structurer et ouvrir les données de la recherche.** La science ouverte porte l'objectif de rendre les données "aussi ouvert que possible, aussi fermé que nécessaire"...

Un espace CY a été ouvert dans l'entrepôt [Recherche Data Gov.](#)

- **Ouvrir et promouvoir les codes sources produits par la recherche,** en favorisant notamment les logiciels libres.



Principes FAIR pour les données de recherche :



- Faciles à trouver (Findable) – s’assurer que les données peuvent être repérées à l’aide de leurs métadonnées et d’un DOI.
- Accessibles (Accessible) – privilégier l’usage de logiciels open source et de formats universellement reconnus (csv, pdf, tif...).
- Interopérables (Interoperable) – utiliser des standards et des formats de données compatibles avec différents systèmes (dublin core...).
- Réutilisables (Reusable) – s’assurer que les données puissent être utilisées et exploitées par d’autres personnes en les dotant d’une licence appropriée comme la Licence Ouverte (licence etalab, CC-BY...)

Que sont les données de recherche ? 1/2

Données de la recherche (Research data) :

« ... 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 des résultats de recherche. » (OCDE, 2007)

Les données peuvent être :

- Déjà existantes
- Être créées dans le cadre de la recherche
- Ou les deux

Que sont les données de recherche ? 2/2

Sources des données

- **Données d'observation :**
capturées en temps réel, hors du laboratoire et impossibles à reproduire (ex : images, données d'enquêtes)
- **Données expérimentales :**
obtenues à partir d'équipements de laboratoire et souvent reproductibles mais coûteuses
- **Données de simulation :**
générées par des modèles informatiques ou de simulation, souvent reproductibles (ex : modèles économiques)
- **Données dérivées ou compilées :**
issues du traitement ou de la combinaison de données brutes (ex : fouilles de texte, bases de données compilées)

Forme des données

- Données textuelles : notes de terrain ou de laboratoire, réponses d'enquêtes (numériques)
- **Données numériques :**
tableaux, comptes, mesures
- **Données audiovisuelles**
(images, sons, vidéos)
- Etc...

Régime juridique

- **Données à caractère personnel** (Règlement général pour la protection des données)
- **Couvertes par un secret** (professionnel, défense ou industriel par exemple)
- Soumises à un **accord de confidentialité**
- Encadrées par des **obligations contractuelles** (contrat d'accès)

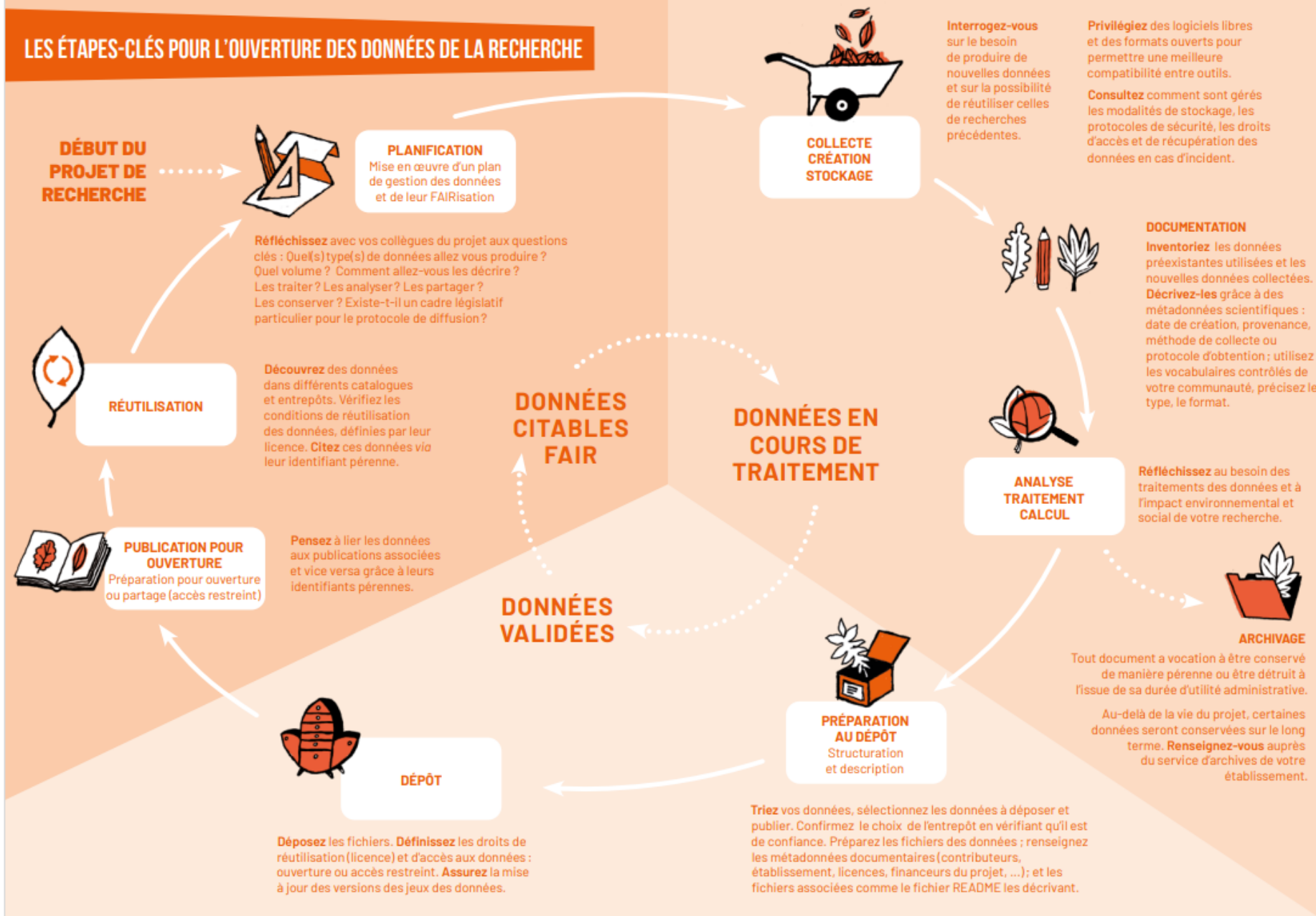
Contexte de production

- Partenariat industriel
- Laboratoire en zone à régime restrictif

Criticité

- Données sensibles
- Confidentielles.

LES ÉTAPES-CLÉS POUR L'OUVERTURE DES DONNÉES DE LA RECHERCHE



Le plan de gestion de données 1/3

Un DMP (data management plan) est un document formalisé qui explicite la manière dont sont obtenues et traitées les données tout au long de leur cycle de vie, de leur collecte à l'archivage. Il doit indiquer :

- quel est le traitement des données de recherche avant, pendant et après la fin du projet,
- les données qui seront collectées, traitées et/ou générées,
- si les données sont partagées, rendues accessibles, comment les données seront organisées et conservées (y compris après la fin d'un projet).

Le plan de gestion de données 2/3

Tous les financeurs publics de la recherche demandent désormais des Plans de Gestion de données.

- « Afin de préparer les données à leur partage et diffusion éventuelle dans le respect du principe « aussi ouvert que possible, aussi fermé que nécessaire » chaque agence demande un plan de gestion des données (PGD) dès le démarrage du projet de recherche et préconise sa mise à jour régulière notamment en fin de projet ». ([Politique de science ouverte des agences de financements françaises, 2022](#))
- Modèles de PGD disponibles et outil de rédaction par financeur disponibles sur [DMP Opidor](#)

Le plan de gestion de données 3/3

Les financeurs exigent 3 versions du même PGD.

Une première version du PGD est attendue à T0+6 mois après le démarrage scientifique du projet. Cette version sera mise à jour à mi-parcours (pour les projets de plus de 30 mois) ainsi qu'à la fin du projet. La première version permet de s'interroger sur le type de données qui sera produite au cours du projet.

Un livrable par projet avec des mises à jour (en moyenne 3 versions selon la durée du projet).

Les parties du PGD (modèle ANR)

1. Description des données et collecte ou réutilisation de données existantes
2. Documentation et qualité des données
3. Stockage et sauvegarde pendant le processus de recherche
4. Exigences légales et éthiques, codes de conduite
5. Partage des données et conservation à long terme
6. Responsabilités et ressources en matière de gestion des données

1. Description des données et collecte ou réutilisation des données existantes

a. Origine des données

La / les source(s) des données : d'où proviennent les données (est-ce que ce sont des données existantes, nouvelles, ou les deux) ?

La méthodologie adoptée pour recueillir les jeux de données (comment accéder aux données, quel a été le processus de création pour obtenir de nouvelles données ? Quels logiciels ont été utilisés pour des nouvelles données ?)

Existe-il des restrictions à la réutilisation des données préexistantes ?



b. Description formelle et technique des données

Quels sont les types de données (la source : données d'observation, expérimentales, de simulation ; la forme : données textuelles, images, audios etc. ; la ou les version(s), le volume : exprimé en espace de stockage ou en quantités) ?

Quels sont les formats des données ? **Privilégier les formats ouverts.**

Justifier le choix de ces formats (s'agit-il de pratiques dans la discipline, une préférence pour les formats ouverts etc.)

1. Description des données et collecte ou réutilisation des données existantes

	Archivage	Image	Dessin	Audio	Texte	Tableur	Vidéo
Formats fermés							
Formats ouverts							

2. Documentation et qualité des données

a. Description des métadonnées et de la documentation des données

Comment les données seront décrites ?

Comment les métadonnées vont être créées et enrichies ?

Indiquer les procédures, les outils ou logiciels utilisés pour recueillir ou éditer les métadonnées

Indiquer les standards de métadonnées (Ex. de standard : Dublin Core, TEI, DDI etc.)

Pensez à un fichier texte "lisez-moi".

b. Mesures de contrôle et de qualité des données

Indiquer les mesures de contrôle et de qualité des données

3. Stockage et sauvegarde pendant le processus de recherche

a. Comment les données et métadonnées seront-elles stockées et sauvegardées durant le processus de recherche ?

Multiplier les différents types de stockage : conserver 3 copies d'une même donnée sur au moins 2 supports différents avec une copie à distance.

Exemples de types de stockage : local, sur ordinateur ; sur un serveur institutionnel ; plateformes externe (HUMA-NUM, Share-Docs, ODS My Core pour le personnel rattaché au CNRS).

Pour les projets CY Générations (ou plus largement tout projet financé par CY) : utiliser le serveur institutionnel de CY pendant le processus de recherche puis déposer ses données dans un entrepôt.

b. Comment la sécurité des données et la protection des données seront-elles assurées tout au long du processus de recherche ?

Qui aura accès aux données (membres du projet ?)

Les données sont-elles sensibles et comment seront-elles gérées ?

Les données sont-elles accessibles par mot de passe ?

4. Exigences légales et éthiques, codes de conduite

a. Si des données à caractère personnel sont traitées, comment le respect des dispositions de la législation sur les données à caractère personnel et sur la sécurité des données sera-t-il assuré ?

Indiquez s'il sera envisagé d'attribuer aux données un identifiant pérenne.

b. Comment les autres questions juridiques, comme la titularité ou les droits de propriété intellectuelle sur les données, seront-elles abordées ? Quelle est la législation applicable en la matière ?

Expliquez quelles conditions d'accès s'appliqueront aux données.

c. Comment les éventuelles questions éthiques seront-elles prises en compte, et les codes déontologiques respectés ?

Déterminez si les questions d'éthique auront une incidence sur la façon dont les données seront stockées et transférées.

[Le comité d'éthique de la recherche de CY Cergy Paris Université \(CER-CY\)](#)

Pratiques éthiques et juridiques

- Respecter les principes directeurs d'une recherche responsable : confidentialité des données sur le terrain et durant le processus de recherche.
- Anonymiser ou pseudonymiser les données personnelles (RGPD). Les données devront dans le cas contraire être détruites.
- Respecter les codes déontologiques en vigueur en Europe.

Respecter le principe « aussi ouvert que possible, aussi fermé que nécessaire ». Toutes les données n'ont pas vocation à être diffusées.

Concilier réglementation sur les données personnelles et recherche

- Entrée en application en mai 2018 du règlement européen sur la protection des données personnelles.
- L'article 89 du RGPD prend en compte les spécificités de l'activité scientifique : réutilisation possible des données à des fins de recherche, traitement de données sensibles (données de santé, sur les appartenances syndicales, les origines ethniques par exemple) à des fins de recherche en prenant des précautions adaptées, dérogations possibles sous certaines conditions à l'obligation d'information des personnes ...

Qu'est-ce qu'une **donnée à caractère personnel** ?

Ce sont **toutes les informations qui permettent d'identifier directement ou indirectement la personne** (Article 4 du RGPD) :

- Les données directement identifiantes : nom, prénom, adresse, photo, voix, etc.
- Les données indirectement identifiantes : un numéro de téléphone, le croisement d'informations, etc.

Qu'est-ce qu'une **donnée sensible** ?

Ce sont les **données qui révèlent la prétendue origine raciale ou l'origine ethnique, les opinions politiques, les convictions philosophiques ou religieuses, l'appartenance syndicale, l'orientation sexuelle, les données de santé, les données biométriques qui permettent d'identifier une personne, les données génétiques.**

L'utilisation de ces données sensibles est possible pour des finalités de recherche publique. Dans tous les cas, une sécurisation renforcée des données et du traitement est nécessaire. Notamment lorsque le consentement des personnes pour traiter leurs données sensibles ne peut être obtenu, un avis préalable de la CNIL doit être demandé avant la réalisation du traitement.

Le traitement des données

- Les **données identifiantes** portant sur un faible nombre de personnes :

Il est important de supprimer le caractère identifiant pour l'étape de publication et/ou à la fin du projet de recherche. Selon les situations, l'anonymisation ou la pseudonymisation sont requises.

- Les **données anonymisées** :

Le lien avec les données personnelles est rompu de manière irréversible. Si l'identification d'une personne n'est possible d'aucune manière, la réglementation sur la protection des données personnelles ne s'applique pas.

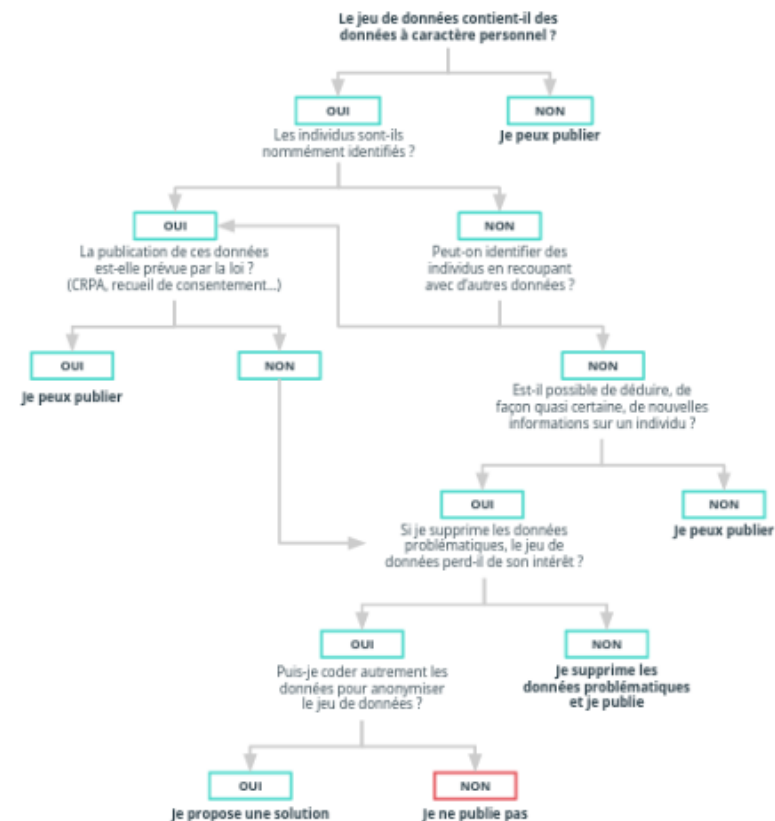
Dans le cadre des enquêtes qualitatives, l'anonymisation ne sera généralement pas possible, dans la mesure où il existe un besoin avéré d'identifier des personnes. Lorsque les objectifs de la recherche nécessitent de mentionner l'identité de l'interviewé (personnalité, expert,...), il convient de leur préciser que des données identifiantes seront publiées et de leur garantir l'accès à la retranscription.

Pour des études quantitatives, l'anonymisation est souvent possible et en adéquation avec le projet de recherche. Il est souhaitable de procéder à l'anonymisation dès la phase de collecte de données.

Exemple d'outils d'anonymisation : [ARX Data Anonymization Tool](#), [Amnesia](#), [μ-ARGUS](#), [scdMicro](#), [Anonimatron](#).

- Les **données pseudonymisées** :

Elles consistent à séparer les données directement identifiantes (exemple : nom et prénom) des autres données non identifiantes (exemple, en attribuant un numéro aux personnes évitant de faire apparaître leur nom, mais en conservant une table de correspondance permettant de remonter à l'identité de la personne).



https://dataactivist.coop/oracle/comprendre_open_data/comprenod.html#47

L'**anonymisation** consiste à rendre impossible toute identification de la personne.

- la **randomisation** : modifier les attributs de telle sorte qu'elles soient moins précises, tout en conservant la répartition globale (ex : permuter les dates de naissance des individus). Il convient de le mentionner dans la description des données.
- La **généralisation** : modifier l'échelle des attributs ou leur ordre de grandeur, afin de s'assurer qu'ils soient communs à un ensemble de personnes (ex : agréger les adresses à l'échelle de l'EPCI).

La **pseudonymisation** est un traitement de données réalisé de manière à ce qu'on ne puisse plus attribuer les données relatives à une personne physique sans information supplémentaire (ex : utilisation d'alias, de numéro séquentiel).

Ces traitements de données sont recommandés par le RGPD pour exploiter des données personnelles dans le respect des droits et libertés des personnes.

Source : CNIL <https://www.cnil.fr/fr/anonymisation-de-donnees-personnelles>

5. Partage des données et conservation à long terme

a. Comment et quand les données seront-elles partagées ? Y-a-t-il des restrictions au partage des données ou des raisons de définir un embargo ?

Comment les données seront rendues disponibles ?

Justifier les restrictions s'il y en a et leur donner une temporalité en les justifiant.

Indiquer qui pourra utiliser les données.

b. Comment les données à conserver seront-elles sélectionnées et où seront-elles préservées sur le long terme (par ex. un entrepôt de données ou une archive) ?

Indiquer quelles données vont être gardées ou contraire détruites (raisons contractuelles, légales, ou réglementaires).

Décrire les utilisations (et/ou les utilisateurs) prévisibles des données dans un cadre de recherche.

Indiquer où les données seront déposées.

c. Quelles méthodes ou quels outils logiciels seront nécessaires pour accéder et utiliser les données ?

Indiquer si les utilisateurs potentiels auront besoin d'outils spécifiques pour l'accès et la (ré)utilisation des données

d. Comment l'attribution d'un identifiant unique et pérenne (comme le DOI) sera-t-elle assurée pour chaque jeu de données ?

Les identifiants pérennes devraient être appliqués de manière à ce que les données puissent être localisées et référencées de façon fiable et efficace.

Indiquer s'il sera envisagé d'attribuer aux données un identifiant pérenne. Typiquement, un entrepôt pérenne de confiance attribuera des identifiants pérennes.

Où déposer ses données ? 1/2

Le choix de l'entrepôt dépend de la nature des données, du projet de recherche, de la pratique disciplinaire. Il existe des entrepôts thématiques, institutionnels, multidisciplinaires.

Re3Data répertorie les entrepôts de données de recherche

Il est recommandé de déposer ses données avant de publier son article afin de lier ses données à l'article en mentionnant le DOI des données.

L'entrepôt devra offrir :

- L'assignation d'un identifiant pérenne type DOI (Principes FAIR = 1 DOI = 1 seul dépôt dans un entrepôt)
- La description des données à un niveau suffisant (métadonnées descriptives standardisées, vocabulaires disciplinaires)
- L'utilisation de licences et des règles d'accès définis dans un cadre légal (licence ouverte)
- Une durée de conservation minimale de plusieurs années

Certains entrepôts dits de confiance sont labellisés par Core trust seal en fonction d'une quinzaine de critères.

Où déposer ses données ? 2/2

Vous pouvez déposer dans :

- **Un entrepôt thématique :**

Par exemple :

[Nakala](#) est un entrepôt de données de recherche pour les Sciences Humaines et Sociales géré par [Huma-Num](#).

[Quetelet-Progado-Diffusion](#) constitue l'entrepôt de confiance pour les bases de données et les enquêtes quantitatives en sciences humaines et sociales.



OU

- **Un entrepôt institutionnel / multidisciplinaire :**

L'université dispose d'un espace institutionnel sur l'entrepôt national [Recherche Data Gouv](#).

Les données peuvent également être déposées dans un entrepôt généraliste, comme [Zenodo](#).

Liste d'entrepôts de confiance : <https://www.ouvrirlascience.fr/donnees-de-la-recherche-comment-choisir-un-entrepot-thematique-de-confiance/>

The logo for the Data repository of CY Cergy Paris Université, featuring the university's logo on the left and the text "Data repository CY Cergy Paris Université" and "(CY Cergy Paris université)" on the right.

UNIVERSITÉ Data repository CY Cergy Paris Université
(CY Cergy Paris université)



6. Responsabilités et ressources en matière de données

a. Qui (par exemple rôle, position et institution de rattachement) sera responsable de la gestion des données (c'est-à-dire le gestionnaire des données) ?

Décrire les rôles et les responsabilités concernant la saisie des données, production des métadonnées, qualité des données, stockage et sauvegarde, archivage et partage des données.

Nommer la(es) personne(s) responsable(s) impliquée(s) dans la mesure du possible.

Indiquer qui est responsable de la mise en œuvre du PGD, et qui s'assure qu'il est examiné et, si nécessaire, révisé.

Envisager des mises à jour régulières du PGD.

b. Quelles seront les ressources (budget et temps alloués) dédiées à la gestion des données permettant de s'assurer que les données seront FAIR (Facile à trouver, Accessible, Interopérable, Réutilisable) ?

Il peut s'agir de frais de stockage, de coût matériel, de temps de personnel, de coûts de préparation des données pour le dépôt, de frais d'entrepôt et d'archivage.

Indiquer si des ressources supplémentaires sont nécessaires pour préparer les données en vue de leur dépôt ou pour payer tous les frais demandés par les entrepôts de données. Si oui, précisez le montant et comment ces coûts seront couverts.

- Service de relecture de PGD de la BU : donnees-bu@ml.u-cergy.fr
- Outil d'aide à la rédaction du PGD avec recommandations : <https://dmp.opidor.fr/>
- [Software Heritage](#) : l'archive ouverte des codes source de logiciels
- [Guidelines for Data Management Plan within the framework of Horizon Europe: RIA, IA, CSA, EIC Pathfinder, Cofund by Université Paris-Saclay](#)

Avant le début de recherche :

- Réfléchir aux données déjà produites ou réutilisées et leur accès
- Anticiper les questions des données personnelles (formulaire de consentement, anonymisation, pseudonymisation)
- Commencer à rédiger le PGD

Pendant le processus de recherche :

- Rédiger le PGD
- Adopter un système d'organisation, de stockage et de nommage des données. Privilégier des formats ouverts. Enrichir les données avec des métadonnées descriptives
- Différencier les données à partager seulement pendant le projet, à partager/stocker plus longtemps, à stocker de façon pérenne
- Multiplier les lieux de stockage : local, sur ordinateurs, sur des serveurs institutionnels, sur des plateformes externes (*HUMANUM*, *Share-docs*, *ODS My Core* pour personnel rattachés au CNRS) etc.
- Anticiper le dépôt des données avant la publication de l'article
- Attribuer un DOI à ses publications, jeux de données, logiciels... HAL ne permet d'attribuer de DOI. Il est possible d'utiliser Zenodo

- DMP Opidor est un outil développé par le CNRS
- Il vous permet de rédiger votre PGD directement en ligne à partir des modèles ANR, Horizon Europe et autres instances de financement
- Vous trouverez des exemples de PGD déjà rédigés
- Une aide est disponible (possibilité de faire appel à la BU pour la relecture de son PGD)
- Des recommandations (en français et en anglais) pour chaque partie sont également disponibles
- Vous pouvez partager et modifier votre PGD entre chercheurs

- Modèles de PGD ANR : [français](#) – [anglais](#)
- Modèles de PGD CNRS : [français](#) – [anglais](#)
- Modèles de PGD Science Europe : [français](#) – [anglais](#)

L'OUTIL OPIDOR

- Accès possible via le site web de la BU :

Données de la recherche

Le service données de la bibliothèque universitaire vous accompagne dans la gestion de vos données de recherche.

- **Comment rédiger son plan de gestion de données (data management plan) :**

Le plan de gestion de données (ou data management plan) est un document de gestion, de synthèse et de pilotage. Il permet d'organiser et d'anticiper toutes les étapes du cycle de vie des données dans un projet. De plus en plus de financeurs demandent un plan de gestion de données dès les débuts du projet dont l'ANR, et la Commission européenne pour Horizon Europe.

La bibliothèque met à disposition une série de recommandations sur [DMPOpidor](#).

- [Baromètre de la science ouverte](#)
- [Données de la recherche](#)
- [Identifiants numériques](#)
- [Portail HAL-CYU](#)
- [Science ouverte](#)

L'OUTIL OPIDOR

1. Se connecter

FR ▼ MON ESPACE

Accès institutionnel

Se connecter

Accès individuel

Pas encore inscrit ? [Créer mon compte](#)

* Courriel


* Mot de passe

Se souvenir de moi

Se connecter

Mot de passe oublié ?

2. S'authentifier via l'accès CY Cergy Paris Université

 Sélectionnez votre établissement

Pour accéder au service DMP opidor sélectionnez ou cherchez l'établissement auquel vous appartenez.

CY Cergy Paris Université x ▼

Sélection

- Se souvenir de mon choix pour cette session.
- Se souvenir de mon choix définitivement et contourner cette étape à partir de maintenant.

 Service Central d'Authentification
Université de Cergy-Pontoise

Identifiant:

Mot de passe:

Prévenez-moi avant d'accéder à d'autres services.

Pour des raisons de sécurité, veuillez-vous déconnecter et fermer votre navigateur lorsque vous avez fini d'accéder aux services authentifiés.

SE CONNECTER EFFACER

Mot de passe oublié ?

Je suis **étudiant**.
Rendez-vous sur la Foire Aux Questions de [MonUCP](#) ("J'ai perdu mon mot de passe. Comment faire?")

Je suis [personnel administratif ou enseignant](#).

L'OUTIL OPIDOR

3. Créer un PGD en fonction des demandes de son financeur



Il faut choisir le modèle de votre financeur.
Vous pouvez choisir la langue de votre choix : français ou anglais.

Créer un plan

1 — 2 — 3 — 4

Choix de la modalité Choix du contexte Choix de la langue Choix du modèle

Sélectionnez la modalité avec laquelle vous souhaitez créer votre plan

Créer un nouveau plan
Vous créez un nouveau plan à partir d'un modèle de plan proposé dans DMP OPIDoR.

Importer un plan existant (Uniquement disponible pour les plans de gestion structurés)

Vous pouvez ensuite remplir les Informations générales et les informations concernant les contributeurs et aller dans *Rédiger*.

L'OUTIL OPIDOR



Cliquez ici pour sélectionner les recommandations appliquées à votre plan 

Vous trouverez ci-dessous la liste des organismes proposant des recommandations et des conseils pour vous guider dans la rédaction de votre plan tout en respectant leur politique de gestion de données.

Vous avez la possibilité de sélectionner 6 organismes au maximum. Puis cliquez pour enregistrer votre sélection.

CY Cergy Paris Université

CY Cergy Paris University (EN)

CNRS

CNRS general (EN)

Data Protection Office - CNRS (EN)

INRIA - Institut National de Recherche en Informatique et Automatique

Inria general (EN)

Institut Pasteur

Institut Pasteur (EN)

Université d'Aix-Marseille

Université d'Aix-Marseille-ENG

Université de Montpellier

Guidance_UM_ENG

Il est possible d'afficher des recommandations personnalisées (en cochant / décochant).

L'OUTIL OPIDOR

4. Remplir son PGD en ligne



Plusieurs onglets facilitent le remplissage du PGD.

L'onglet Partager permet d'inviter d'autres collaborateurs à alimenter le plan de gestion des données.

L'onglet Demande d'assistance conseil permet de nous saisir d'une question ou d'un problème.




A la fin du remplissage, vous pouvez ensuite cliquer sur Télécharger et choisir le format de votre choix.

L'OUTIL OPIDOR

L'outil DMP Opidor permet également d'avoir accès à des exemples de DMP déjà rédigés. C'est l'accès DMP Publics.

PGD publics

Les PGD publics sont créés à l'aide de DMP OPIDoR et partagés publiquement par leurs propriétaires. Ils n'ont pas été vérifiés pour leur qualité, leur exhaustivité ou leur adhésion aux lignes directrices des financeurs.

Titre du plan	Modèle	Domaines de recherche	Organisme	Dernière mise à jour	Télécharger
"Reproducibility with VIP" project DMP	ANR - DMP template (english)	Health sciences; Computer and information sciences	CNRS	15/11/2024	
"THEIA Data center of Montpellier" entity DMP	AgroParisTech - DMP template for Entity		INRAE - Institut national de recherche pour l'agriculture l'alimentation et l'environnement	12/11/2024	
DMP du projet "Valeur du temps : Explorer les préférences individuelles face à des temps d'attente incertains"	Science Europe : modèle structuré	Economics and business	Université de Rennes	12/11/2024	
CDS-THEIA-MTP project DMP	Science Europe: structured template for research entities	1.5 Earth and related environmental sciences	INRAE - Institut national de recherche pour l'agriculture l'alimentation et	08/11/2024	