

FICHE PRATIQUE : faire un PLAN de GESTION de DONNÉES (PGD ou DMP) à l'échelle d'une structure (établissement, laboratoire, équipe de recherche)

3 QUESTIONS pour éclairer le sujet :

Qu'est-ce qu'un plan de gestion de données « structure » ?



C'est un **document de référence pour la structure** (établissement, laboratoire, équipe) sur la gestion des données qu'elle produit dans le cadre de son activité de recherche.
Il est élaboré de manière collective, dans une démarche de **définition d'une politique de gestion des données de recherche** pour la structure.
Le « PGD structure » a un périmètre plus large que le « PGD projet », une durée de vie plus longue.

Que disent les textes officiels ?

” Les établissements publics et fondations reconnues d'utilité publique [...] **définissent une politique de conservation, de communication et de réutilisation des résultats bruts des travaux scientifiques** menés en son sein. A cet effet, ils veillent à la **mise en œuvre par leur personnel de plans de gestion de données** et contribuent aux infrastructures qui permettent la **conservation, la communication et la réutilisation des données et des codes sources**. ”

[Décret n°2021-1572](#) du 03/12/2021 relatif au respect des exigences de l'intégrité scientifique, article 6

Pourquoi élaborer un plan de gestion de données structure ?



Pour **harmoniser les pratiques de gestion/ouverture des données au sein de la structure**



Pour **monter en qualité** sur la gestion des données



Pour **faciliter la rédaction des « PGD projets »** pour les chercheur·se·s et personnels de recherche, et **respecter les exigences des financeurs** de la recherche publique



Pour **préserver le patrimoine scientifique** de la structure :

- **conserver et valoriser les données produites au cours du temps**
- **documenter** les données produites
- **ne pas « perdre »** les données produites par les **doctorant·e·s** à leur départ



Pour adopter une **démarche « science ouverte »** et respecter les **principes FAIR**



Pour **faciliter la réutilisation des données de recherche** : le 1er ré-utilisateur potentiel est **le·la chercheur·se** à l'origine des données, ou bien ses collaborateur·rice·s

Par où COMMENCER ?

Identifier les données générées par la structure de recherche...

- Données d'**observation** ? (capturées en temps réel : imagerie médicale, données d'enquête, relevés de capteurs...)
- Données **expérimentales** ? (obtenues à partir d'équipements de laboratoire : chromatogrammes, séquençage ADN...)
- Données **computationnelles** ? (générées par des modèles informatiques ou de simulation)
- Données **dérivées** ou **compilées** ? (issues du traitement de données brutes : fouille de texte, bases de données compilées...)
- Données **de référence** ? (collection de jeux de données revus par les pairs et mis à disposition : GenBank, base de données de l'IGN, archives d'images historiques...)



... et en déduire les éventuels points de vigilance :

- **Volumétrie** importante ?
- **Accès** fréquents et/ou traitements intensifs sur les données ?
- Données à **caractère personnel** ? Voire données sensibles ? (*données de santé, relatives à l'orientation sexuelle, politique ou syndicale, relatives à l'appartenance à un groupe ethnique, relatives à des infractions ou condamnations judiciaires*)
- Données qui relèvent de la **PPST** ? (protection du potentiel scientifique et technique de la nation)
- Données qui peuvent donner lieu à des **innovations**, des **brevets** ? Générées dans le cadre de partenariats avec des entreprises privées ?
- **Réutilisation** de données produites par d'autres, donc éventuels problèmes de droits d'auteur ou de propriété intellectuelle ?
- Données qui mettent en jeu le **droit à l'image des personnes** ?



Choisir un mode d'organisation pour le PGD :

- Par type d'activités de la structure ?
- Par type de données produites ?
- Selon le mode d'obtention, la nature ou le format des données ?

Exemple : les données produites en ZRR ne relèvent pas toutes de la PPST, elles peuvent donc faire l'objet de gestion différenciée

Pour vous aider à élaborer un plan de gestion des données de recherche produites par votre structure, nous vous proposons une trame, à ajuster en fonction de vos besoins, spécificités et priorités :

1- INFORMATIONS sur la STRUCTURE

- > Acronyme, nom, **identifiants** (n°[RNSR](#), [idHAL](#), [ROR](#)...)
Présenter succinctement la structure de recherche, les champ(s) disciplinaire(s)
- > Identifier les **personnes référentes** (nom et [ORCID](#)), par exemple :
 - responsable de la structure
 - responsable Science ouverte ou gestion des données
 - responsable informatique
- > Identifier les **partenaires** impliqués dans la structure
- > Lister les **documents de référence** qui encadrent l'activité de la structure, par exemple la charte ou feuille de route Science ouverte de l'établissement, le règlement de ZRR etc.

2- ORIGINE des DONNÉES PRODUITES ou RÉUTILISÉES

- > Identifier les **données produites** : nature, type, volume, formats, etc.
- > Identifier les **données utilisées** mais provenant de sources extérieures
→ Donner les sources + conditions d'utilisation (droits)
- > Identifier d'éventuels **autres produits de recherche** : logiciels et codes sources, workflows, modèles, échantillons physiques, etc.
- > Si justifié : décrire les **outils spécifiques** ou **plateformes technologiques** utilisées pour la collecte et/ou le traitement des données

3- DOCUMENTATION des DONNÉES

- > Définir/choisir les **métadonnées** utilisées a minima pour décrire les données produites (ex : identité du producteur/de la productrice des données, date de collecte, outil et unité de mesure, etc)
Éventuellement, choisir un **standard de métadonnées** (ex : EML en écologie, DDI pour les sciences sociales)
- > Choisir des **référentiels reconnus dans la communauté** pour décrire et caractériser les données : thésaurus, ontologie, référentiel taxonomique, vocabulaire contrôlé, etc.



A titre d'exemple, voir la liste des [référentiels « mots clés » utilisés dans le catalogue de métadonnées dat@UBFC](#)

- > Définir une **convention de nommage de fichiers**. Avoir des règles communes permet d'identifier et de partager plus facilement les fichiers, mais aussi de garder la maîtrise de leur contenu.
- > Lister les éléments retenus pour **documenter les données** (tout ce qui peut permettre de garder la maîtrise du contenu, de comprendre, lire, interpréter, conserver, réutiliser les données) :
 - Dictionnaire de données
 - Définition des variables
 - Protocole d'acquisition des données
 - Fichier README
 - Cahiers de laboratoire
 - etc.

4- STOCKAGE INTERMÉDIAIRE des DONNÉES



STOCKAGE : enregistrement d'une information sur un support physique – le support physique peut avoir des caractéristiques très variées.

Le **stockage intermédiaire** concerne le stockage des données actives, de travail (=données dites « chaudes » ou « tièdes ») contrairement aux données finalisées, sur lesquelles on ne travaille plus (=données « froides »).

4- STOCKAGE INTERMÉDIAIRE des DONNÉES (suite)

> Expliquer quelle **infrastructure** au sein de la structure de recherche permet de stocker les données intermédiaires (serveur par exemple, solution type Next Cloud, etc)

> Expliquer quelles **stratégies de sécurité** sont en place :

- Contrôle des accès (physiques et distants)
- Sauvegardes

> Définir des **recommandations** en termes de stockage intermédiaire et de sauvegarde



RECOMMANDATIONS concernant l'ESR : le recours aux clouds privés (AWS, Google Drive, Drop Box, iCloud, etc) est fortement déconseillé. Seuls 5 prestataires privés sont qualifiés par l'ANSSI ([voir ce document, page 13](#)). Dans le cas des ZRR, le recours aux clouds privés est strictement interdit.

5- GESTION des DONNÉES de THÈSE

> Dispositions particulières pour ces données ?

6- CADRE LÉGAL, RÉGLEMENTAIRE, ÉTHIQUE

> Titularité des **droits sur les données** : les données de recherche appartiennent au laboratoire, contrairement aux publications ; mais cas particuliers des bases de données et des codes sources

> **Partenariat** avec des entreprises ou partenariats internationaux qui ont des implications en termes de droits ?

> Indiquer si la structure collecte ou produit des données qui **relèvent d'une obligation particulière** et expliquer les aménagements mis en place. Par exemple :

- **Données à caractère personnel** (application du RGPD)
- **Données de santé, RIPH** (Recherche Impliquant la Personne Humaine)
- Données issues de **ressources génétiques** (protocole de Kyoto)
- Données qui doivent rester **confidentielles** (données qui relèvent de la PPST, innovations/brevets, intelligence économique, secret bancaire, secret de l'instruction, etc)

> Si un cadre **éthique** ou **déontologique** existe dans le domaine de recherche, y faire référence

> Si une **démarche qualité** existe dans la structure, y faire référence

7- OUVERTURE des DONNÉES

> Définir la **politique d'ouverture des données** :

- Quelles données ouvrir ?
- Quand ouvrir les données (définition d'un embargo par exemple) ?
- Ouvrir les données qui viennent en supplément d'un article ?



Il est recommandé de **ne pas déposer les données chez un éditeur privé en « supplementary material »**, car dans ce cas, les données ne sont pas systématiquement libres d'accès ni correctement documentées.

> Préciser le(s) **entrepôts** choisi(s) par la structure pour l'ouverture des données



RECOMMANDATIONS pour le choix d'un entrepôt :

1. Privilégier un entrepôt **disciplinaire** (reconnu dans la discipline)
2. Privilégier un entrepôt **de confiance** (qui offre un service de qualité en termes de sécurité et de pérennité des accès), par exemple une solution **institutionnelle**
3. Si pas d'entrepôt disciplinaire et/ou institutionnel, aller vers un entrepôt **généraliste de confiance**, comme Recherche Data Gouv par exemple ou l'entrepôt **recommandé par le financeur public**
4. Choisir un entrepôt qui permet d'**attribuer des identifiants pérennes** type DOI et des **licences** pour encadrer la réutilisation des données

7- OUVERTURE des DONNÉES (suite)

- > Si publication des données souhaitée via un **data paper**, indiquer la revue choisie
- > **Description/recensement** des données dans le [catalogue dat@UBFC](#) ?
(cela peut permettre de donner de la visibilité et un DOI à des données qui ne peuvent pas être ouvertes)

8- CONSERVATION et ARCHIVAGE des DONNÉES



ARCHIVAGE = ensemble d'actions qui a pour but de garantir sur le long terme l'accessibilité à des informations (dossier, documents, données) que l'on doit ou souhaite conserver pour des raisons juridiques, historiques ou culturelles (*Wikipedia*)

Dans le cas de l'archivage à long terme de contenus numériques, il faut assurer :

- La **pérennité** des supports de stockage
- L'**accès au contenu** même quand les formats de données deviennent obsolètes
- L'**intégrité** des données

- > Déterminer quelles données doivent être conservées, combien de temps et sur quel(s) support(s)
- > Déterminer quelles données doivent être détruites, et à quelle échéance



RECOMMANDATIONS : conserver les données nécessaires à la reproductibilité du résultat

- **Données de test** → à supprimer en fin de projet
- **Données de validation** associées à une publication → à garder 5 à 10 ans après la publication
- **Données difficilement reproductibles** ou réutilisables dans d'autres domaines → à garder + de 10 ans après la publication
- **Données non reproductibles** ou à forte valeur → à archiver

L'opérateur agréé pour l'archivage des données de l'ESR est le **CINES** – pour travailler avec le CINES, un contrat doit être passé à l'échelle de la structure de recherche et cela a un coût qu'il faut prévoir.

9- PERSPECTIVES

- > Identifier les **marges d'amélioration**, **planifier des évolutions** ou des **dépenses** (ex : achat de matériel informatique, recrutement d'un stagiaire pour effectuer une mission particulière, etc)



Évaluer les **coûts de gestion des données** : [infographie + modèle de tableau de chiffrage à télécharger – DoRANum](#) → Ces coûts peuvent être inclus dans le montage des projets car les principaux financeurs publics (ANR et Europe notamment) acceptent de **financer les coûts liés à la Science ouverte**.

Des RESSOURCES à votre disposition :



N'hésitez pas à contacter les **documentalistes de l'Atelier de la donnée dat@UBFC** :

- en écrivant à cette adresse dataubfc-doc@ubfc.fr
- ou en remplissant ce **formulaire de contact**
<https://data.ubfc.fr/services/animations/>



Une sélection de **ressources synthétiques et didactiques** sur toutes les étapes du cycle de vie des données de recherche, disponible dans la rubrique « **Pour aller plus loin** » de notre site Web :

<https://data.ubfc.fr/pour-aller-plus-loin/>

SOURCES utilisées pour créer cette fiche :

- ✂ **Modèle de PGD structure d'Agro Paris Tech**, disponible sur DMP OPIDoR
https://dmp.opidor.fr/template_export/2059692581.pdf
- ✂ **PGD structure – Atelier distant sur la formation aux PGD** (25 juin 2020) – Dominique L'HOSTIS et Sylvie COCAUD – INRAE DIPSO
https://urfirstinfo.hypotheses.org/files/2020/07/PGD-Structure_SC-DLH_GTDMP_25062020.pdf
- ✂ **Partager les données liées aux publications scientifiques – Guide pour les chercheurs** – CoSO
https://www.ouvrirscience.fr/wp-content/uploads/2022/04/Guide-Partager_les_donnees_pour_impression.pdf
- ✂ **Séminaire stockage – Cellule Data UGA/INIST-CNRS/URFIST**
Lyon – 25 mai 2021
https://doranum.fr/stockage-archivage/seminaire-stockage-des-donnees-de-la-recherche_10_13143_kp98-bj30/



Source des pictogrammes : <https://iconmonstr.com/>