CELLULE DATA
GRENOBLE ALPES

Rédiger un Plan de Gestion des Données

Séminaire CDGA, 30 avril 2021

CELLULE DATA
GRENOBLE ALPES

1. Contexte, enjeux et cadre légal
2. Le plan de gestion des données (PGD)
3. Exemple du PGD de l'ANR
4. Aide de la Cellue Data Grenoble Alpes

Contexte, enjeux et cadre légal

CELLULE DATA
GRENOBLE ALPES

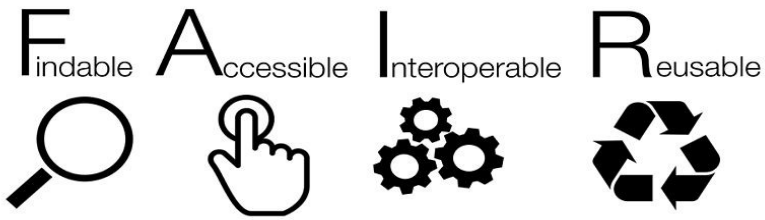
Ouverture des données de la recherche

Par défaut, l'**ouverture des données de la recherche est obligatoire selon le principe "aussi ouvert que possible, aussi fermé que nécessaire"**.

Exceptions :

- Données **personnelles** non anonymisées ou pseudonymisées
- Données **non achevées** (sauf données géographiques qui peuvent être inachevées, directive INSPIRE), certaines données environnementales (cf protection des espèces)
- **Photos** quand il y a une personne reconnaissable
- Données **scrapées** (car souvent interdit par les conditions d'utilisation) sauf :
  - si le préjudice du producteur est nul (risque encouru nul également)
  - le scraping d'une partie non substantielle des données reste possible

Principes FAIR



Findable / *Trouvable*

Données **faciles à trouver**.

- possédant un identifiant unique et pérenne
- décrites par des métadonnées riches
- enregistrées ou indexées dans une source interrogeable

Interoperable / *Interopérable*

Facile à combiner avec d'autres jeux de données, par les humains **et** les systèmes informatiques

- formats libres et ouverts
- mise à disposition du code source si le logiciel de traitement existe
- standards de métadonnées et vocabulaire standardisés

Accessible / *Accessible*

Données ou au moins méta-données facilement accessibles.

- entrepôt de confiance, pérenne, certifié
- définir les conditions d'accès et la licence de diffusion
- si embargo ou accès restreint : méta-données accessibles

Reusable / *Réutilisable*

Prêtes à être **réutilisables** pour une future recherche y compris via des méthodes informatiques

Enjeux des données de la recherche

- En premier lieu **faciliter le travail de recherche** :
  - Assurer la **conservation** des données à moyen terme
  - Pouvoir **reproduire** les résultats scientifiques, y compris par l'équipe qui les a obtenus !
  - Pouvoir facilement les **réutiliser** pour produire de nouvelles recherches
  - **Valoriser** les résultats scientifiques, mais aussi les données et les codes qui ont permis de les obtenir et en augmenter la visibilité
  - Favoriser de **nouvelles collaborations**, de nouvelles approches
- Enjeux d'ordre sociétaux :
  - Assurer une **souveraineté** sur les données produites
  - Assurer l'**intégrité scientifique**
  - Garantir la **transparence**, et assurer la confiance des citoyens en la recherche

Dans un cadre légal de plus en plus structuré

- **RGPD** (2016) : règlement européen sur la protection des données personnelles

- Loi pour une république numérique (2016), Plan National pour la Science Ouverte (2018), H2020 (depuis 2017) et Horizon Europe (à partir de 2021), Feuille de route de la science ouverte du CNRS (2018), Plan données de la recherche CNRS (2020) : l'**ouverture des données de recherche financées sur fonds publics devient la norme**

- **Obligation de Plan de Gestion de Données** exigée par de plus en plus de financeurs de la recherche (H2020, ANR, Horizon Europe, ...)

Un plan de gestion des données

Plan de Gestion de Données (PGD)

Ce n'est pas :

- un nième document administratif …

C'est :

- une **aide concrète à la gestion des données** durant et après la phase de recherche
- un outil normalisé et évolutif tout au long du projet
- un livrable du projet

Objectif : anticiper la gestion des données dans tous les aspects de la recherche en se posant les bonnes questions

Exigé par des financeurs :

- Horizon 2020 depuis 2016
- ERC depuis 2017
- ANR depuis 2019

Plan de Gestion des Données

Le PGD (ou DMP, Data Management Plan) **permet de réfléchir à la gestion des données d'un projet en amont afin de l'anticiper** :

- **Quelles données vont être obtenues :** quels types de données, comment sont-elles collectées, où les stocker, comment on sécurise le stockage, quelle volumétrie, quels formats, quelle organisation ...
- **Comment vont-elles être utilisées :** comment on les partage, comment on les traite, où on les traite, ...
- **Comment elles vont être préservées :** à quel terme, quelles données, où, comment ...
- **Comment elles vont être valorisées :** comment les diffuser, sous quel format, sous quelle licence, quelles données, comment associer les codes, …
- **Comment assurer le financement des ressources nécessaires ?**

Pour répondre aux principes FAIR de l'open data

Plan de Gestion des Données
En pratique

Le calendrier pour les ANR et les projets européens :

- 1ere version (obligatoire) : 6 mois après le démarrage du projet
- Mises à jour (recommandées) à chaque évaluation du projet
- Version finale à la fin du projet (évaluation)

A noter : la première version du PGD n'implique pas des réponses complètes et précises à l'ensemble des questions. Elle offre surtout la possibilité pour les porteurs du projet de réfléchir aux différentes problématiques liées à la gestion des données.

Plan de Gestion des Données
Projet ANR

6 parties :

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**

Source :
https://anr.fr/fr/actualites-de-lanr/details/news/lanr-met-en-place-un-plan-de-gestion-des-donnees-pour-les-projets-finances-des-2019/

Plan de Gestion des Données
Projet européen

6 parties :

1. **Présentation des données**
2. **Données FAIR**
   - Rendre les données trouvable, y compris les métadonnées
   - Rendre les données librement accessibles : diffusion des données
   - Rendre les données interopérables
   - Augmenter la réutilisation des données (licences)
3. **Répartition des ressources**
4. **Sécurité des données**
5. **Aspects éthiques**
6. **Autres aspects**

Source :
https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm

Le plan de gestion des données : exemple de l'ANR

CELLULE DATA
GRENOBLE ALPES

Collecter et produire ses données

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

1a. Comment de nouvelles données seront-elles recueillies ou produites et/ou comment des données préexistantes seront-elles réutilisées ?

En cas de réutilisation de données, provenance et mode de collecte
En cas de production, méthodes, logiciels, outils utilisés

1b. Quelles données (types, formats et volumes par ex.) seront collectées ou produites ?

Nature des données produites ou réutilisées (par ex : RMN) ?
Type des données (par exemple image, audio, vidéo, texte, numérique, etc.) ?
Format (ouvert ou fermé)
Volumétrie prévisionnelle

Plan de Gestion des Données
Projet ANR

Exemple (Palagram) :

Ombro_IMFT : Ombroscopie Courant Stationnaire (IMFT)

1a. Comment de nouvelles données seront-elles recueillies ou produites et/ou comment des données préexistantes seront-elles réutilisées ?

Les données seront recueillies par une caméra pco edge scmos (résolution 2400x2000) à fréquence contrôlable (0-100Hz), disponible à l'IMFT. La résolution sera de 1 mm/pixel

Aucune donnée préexistante sur cette configuration.

1b. Quelles données (types, formats et volumes par ex.) seront collectées ou produites ?

Les données sont des images au format .tiff avec un fichier de métadonnée par séquence (au format texte .rec). Le volume estimé est de 6To.

Documenter ses données

2. Documentation et qualité des données

2a. Quelles métadonnées et quelle documentation ?

Organisation des données

Structuration des dossiers)
Nommage des fichiers (ex : Doranum)
Gestion de versions

Description des données

Champs de description (ex : variables, les unités de mesure…)
Standards utilisés (ex DataCite, DublinCore)

2b. Quelles mesures de contrôle de la qualité des données seront mises en œuvre ?

Par exemple, validation de données par l'ensemble des scientifiques impliqués dans le projet.

Plan de Gestion des Données
Projet ANR

2.a Documentation et qualité des données — Exemple (HiTrac, Atlas, IAB) :

Les données seront systématiquement classées par dossiers, un dossier correspondant à une série d'expérience et nommé selon le format : AAAA-MM-DD.description succinte

Dans chaque dossier et dès que ce sera pertinent, un fichier "readme.txt" chapeautera le reste de l'arborescence pour donner les détails des acquisitions effectuées non enregistrés dans les fichiers de métadonnées automatiquement créés par les instruments utilisés.

Au sein de chaque dossier, chaque expérience sera rangée dans un sous-dossier avec une dénomination claire (à adapter en fonction de l'expérience)

Les fichiers analytiques générés à partir de la plateforme de collecte seront documentés selon la spécification DDI (*Data Documentation Initiative* http://www.ddialliance.org/) dédiée à la documentation de données d'enquêtes quantitatives en sciences humaines et sociales. La documentation des variables sera effectuée avec le logiciel *Nesstar Publisher* du *Norwegian Centre for Research Data*. Ce logiciel permet notamment de produire un dictionnaire détaillé des variables et un export des fichiers de données labellisées au format Stata et SPSS. Une copie des données au format CSV (fichiers texte) sera systématiquement conservée.

Les données sont accompagnées des **métadonnées** attendues par l'entrepôt de données NCBI/GEO en vue de leur publication.

Les métadonnées attendues sont décrites ici :
https://www.ncbi.nlm.nih.gov/geo/info/seq.html#metadata

Plan de Gestion des Données
Projet ANR

Gérer ses données pendant le projet

3- Stockage et sauvegarde pendant le processus de recherche

3a. Comment les données et les métadonnées seront-elles stockées et sauvegardées tout au long du processus de recherche ?

Modalités et lieu de stockage (type d'hébergement, support)
Procédure de sauvegarde
Utilisation de plateformes de sauvegarde et de partage institutionnelles (ex MyCore ou Cloud UGA, GRICAD)

3b. Comment la sécurité des données et la protection des données sensibles seront-elles assurées tout au long du processus de recherche ?

Modalités d'accès aux données pendant le projet
Droits d'accès (Pour les membres, les partenaires)

Plan de Gestion des Données
Projet ANR

Gérer ses données pendant le projet

Exemple (IAB) :

Nous définissons ici trois espaces de stockage :

- un espace de stockage **sécurisé**, **privé** (*summer,* données froides et tièdes)
- un espace de stockage destiné au **calcul intensif** (*bettik,* données chaudes et tièdes)
- un espace d'**archivage** pérenne, public (l'entrepôt de données NCBI/GEO, données froides)

Au cours du processus de recherche les données sont stockées sur *summer* et *bettik*.

Les **données brutes** (*fastq,* niveau 1) que nous produisons sont stockées sur notre espace de stockage sécurisé hébergé par la plateforme *summer* de l'unité Grenoble Alpes Recherche - Infrastructure de Calcul Intensif et de Données (GRICAD UMS 3758). Cette solution de stockage et de sauvegarde est réparties dans trois centres de données de l'UGA situés sur le campus de Saint Martin d'Hères. L'infrastructure en place est basée sur les technologies du constructeur NetApp. Nous disposons de manière privative d'une volumétrie de 60 To (évolutif) associée à des fonctions évoluées assurant la sécurité, l'intégrité des données et la continuité de service. Notre espace de stockage sécurisé est directement accessible depuis les serveurs du mésocentre de calcul CIMENT/GRICAD.

Les **données traitées** (niveau 2 et 3) sont stockées sur l'espace destiné au calcul intensif (*bettik)* du mésocentre CIMENT/GRICAD. *Bettik* est un espace de stockage distribué haute performance partagé (*scratch*) accessible depuis les clusters du mésocentre. Il offre une volumétrie conséquente (>1Po) partagée par l'ensemble des noeuds de calcul avec des temps d'accès minimaux et des débits élevés. *Bettik* utilise le système de fichier distribué *BeeGFS*. Les données traitées sont produites par l'exécution de nos pipelines sur les infrastructures de calcul du mésocentre.

Les questions juridiques et éthiques

4. Exigences légales et éthiques, codes de conduite

4a. 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 ?

Formalités de déclaration de données sensibles (par exemple auprès du DPO, etc.) ?
Mesures prises pour garantir la confidentialité des données personnelles (pseudonymisation ou anonymisation, chiffrement, procédure d'accès…) ?
Modalités d'information des personnes (formulaires de consentement, etc.)

4b. 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

Titularité des droits de propriété intellectuelle (données, code, etc.)
Licences d'usage des données ?
Respect des règles des différentes tutelles.

4c. Comment les éventuelles questions éthiques seront-elles prises en compte, les codes déontologiques respectés ?

Les questions juridiques et éthiques

Exemple (Atlas) :

Un consentement éclairé et écrit sera demandé aux agent·e·s dispensateurs sélectionné·e·s pour cette étude. Un numéro de participation sera mentionné sur le formulaire de consentement afin de leur permettre d'exercer leur droit d'accès, de rectification et/ou d'opposition.

Comme mentionné sur les notices d'information, les données brutes seront conservées pour une durée maximum de 5 ans après la fin du projet, soit jusque fin 2026 au plus tard. Seules des données ayant fait l'objet d'une anonymisation feront l'objet d'un archivage de longue durée (voir le chapitre "Sélection et préservation").

L'ensemble des formulaires de consentement et des notices d'informations sont disponibles en annexe du protocole, sur le site https://atlas.solthis.org/

Les formulaires de consentement seront conservés pendant 5 ans après la fin du projet (soit jusque fin 2026) dans une armoire sous clé dans les locaux de Solthis situés à Abidjan pour les formulaires de consentement collectés en Côte d'Ivoire, Bamako pour les formulaires collectés au Mali et Dakar pour les formulaires collectés au Sénégal.

Diffuser ses données

5. Partage des données et conservation à long terme

5a. 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 ?

Mode de diffusion des données :

Libre ? Restreint ? Embargo ?

Quels jeux de données librement accessibles ? (données liées aux publications?)

Lieu de publication et de diffusion des données.

Utilisation d'archives ouvertes comme HAL ou OpenEdition, d'entrepôts de données thématiques ou généralistes ?

Délai de diffusion

5c. Quelles méthodes ou quels outils logiciels seront nécessaires pour accéder et utiliser les données ?

5d. Comment l'attribution d'un identifiant unique et pérenne (comme le DOI) sera-t-elle assurée pour chaque jeu de données ?

Utilisation d'un entrepôt de données pour l'obtention d'un DOI (par ex : Dataverse, Zenodo, ou Nakala)?

Conserver à long terme ses données

5. Partage des données et conservation à long terme

5b. 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) ?

Utilité des données ('data utility') ?

Contexte, public, durée d'utilisabilité, etc.

Quelles données seront conservées sur le long terme ?

Procédures de la conservation à long terme (par ex : entrepôt certifié) ?

Utilisation d'une plateforme d'archivage pérenne (CINES en France) ?

Conditions d'archivage (budget nécessaire)

Responsabilité des données

Plan de Gestion des Données
Projet ANR

Diffuser ses données

Exemple (Atlas, HiTrac) :

Les fichiers de données anonymisées seront déposés sur le dépôt de données Zenodo : https://www.zenodo.org/communities/atlas-research/. Zenodo est un dépôt gratuit et public, hébergé par le CERN et appuyé par l'Union Européenne.

Les données déposées sur Zenodo disposeront chacune d'un DOI (Digital Object Identifier) permettant leur identification pérenne au cours du temps. Ces DOI seront notamment utilisés dans les publications scientifiques générées à partir de ces données.

Pendant la durée du projet, les fichiers de données seront déposés en accès à la demande, le temps que les équipes de recherche publient les résultats de leurs analyses.

Une fois le projet terminé et les articles correspondants publiés, les fichiers seront en accès libre et distribués sous licence *Creative Commons - Attribution - Partage dans les Mêmes Conditions 4.0 International* (https://creativecommons.org/licenses/by-sa/4.0/).

Afin d'éviter tout biais dans la sélection des données, toutes les données exploitables seront conservées. En l'absence de lieu de stockage dédié à l'Université Grenoble Alpes, les bandes seront stockées en doubles exemplaires dans deux laboratoires différents de l'université.

Responsabilité et budget

6. Responsabilités et ressources en matière de gestion des données

6a. Qui (par exemple rôle, position et institution de rattachement) sera responsable de la gestion des données (c'est-à-dire le gestionnaire des données) ?

Responsable de la gestion des données au sein du projet

6b. 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) ?

Le budget (frais de stockage, coût de traitement, coût matériel, d'archivage)
Les ressources humaines (métier)

Plan de Gestion des Données
Intégrer les bonnes pratiques

Pour résumer, les bonnes pratiques tout au long du projet :

- **Documenter les données** : description du projet, du processus de collecte, des matériels et logiciels utilisés, de la structuration de la base de données, du processus de nettoyage, …
- **Utiliser des métadonnées** standards et spécifiques à sa communauté
- **Utiliser des formats de fichiers ouverts**, non propriétaire, documentés, reconnus dans sa communauté (https://facile.cines.fr/)
- **Utiliser des conventions de nommage et d'organisation** des fichiers et répertoires, préciser les versions et dates (ou utiliser un gestionnaire de version)
- **Définir les conditions juridiques d'utilisation** de ces données
- **Définir les modalités de diffusion**, de stockage et d'archivage des données

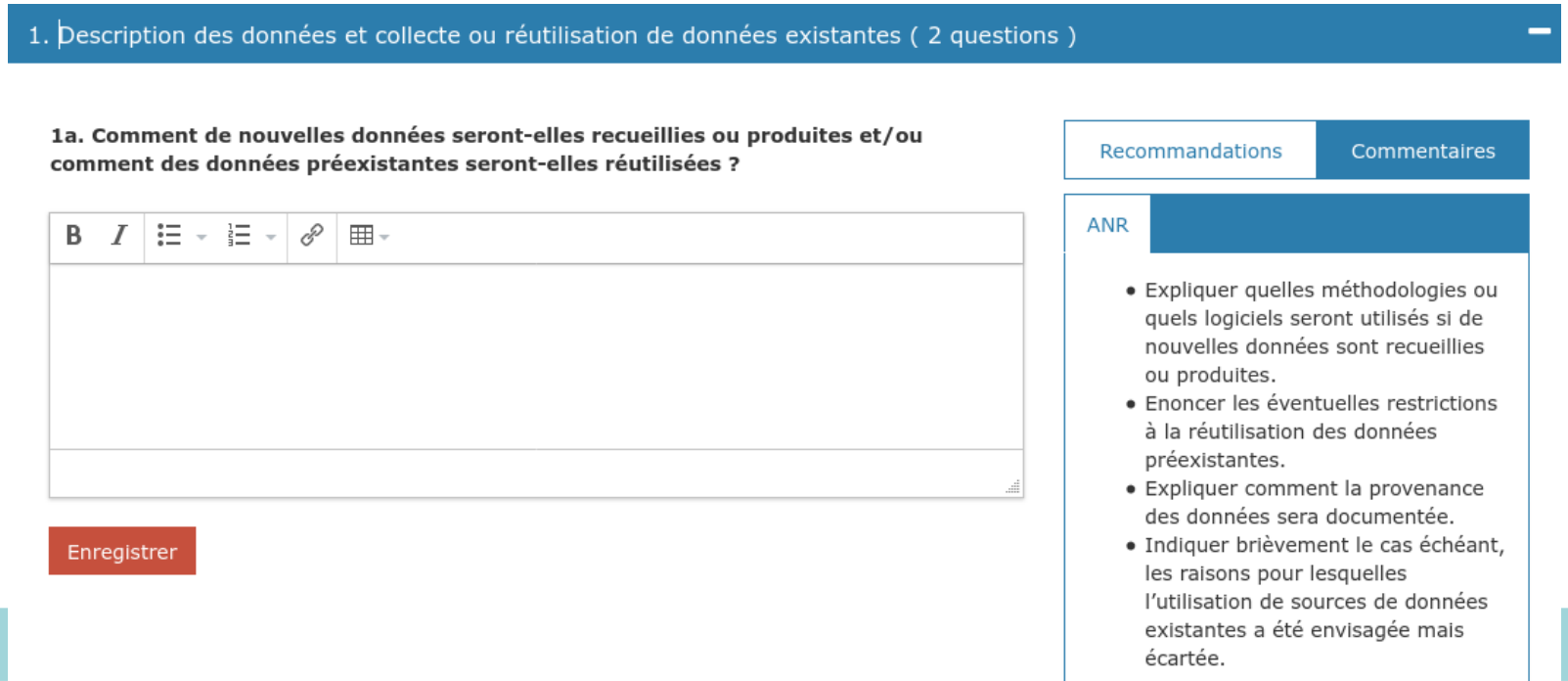
Plan de Gestion des Données
En pratique

L'outil DMP Opidor :

- Adapté aux différents modèles de DMP (modèle des financeurs : ANR, Europe, ...)

- Fonctionnalités utiles :
  - recommandations pour les réponses aux questions
  - demande d'aide intégrée pour solliciter les cellules d'accompagnement comme la Cellule Data Grenoble Alpes
  - consultation des DMP publics

- Travail en cours pour intégrer des recommandations propres au site de Grenoble Alpes



Pour vous aider, une cellule d'accompagnement sur les données de la recherche

CELLULE DATA
GRENOBLE ALPES

Aide sur les PGD

Aide à l'utilisation de l'outil **DMP Opidor.**

Aide à la **rédaction, relecture et commentaires.**

En particulier, élaboration de documents qui regroupent tous les éléments techniques concernant les **plateformes de stockage UGA** (SUMMER, Bettik, Silenus, Mantis...) à intégrer dans le DMP.

Actions :

Prise de contact avec les porteurs de projets ANR 2020.
Communication avec la DGD RIV.

Expérience :

8 accompagnements en cours

Et au-delà ...

Pour rappel :

Répondre à toutes les questions liées aux données

S’appuyer sur les **expertises présentes** sur le site pour répondre aux questionnements des scientifiques

Faire de la **veille juridique et réglementaire**

Aide sur le stockage

Aide sur le traitement des données

Aide sur la diffusion des données et leur description

Contacts

- Pour contacter la cellule :
  uga-cellule-data@univ-grenoble-alpes.fr

- Pour se tenir informer : abonnement sur la liste uga-research-data
  - https://listes.univ-grenoble-alpes.fr/sympa/info/uga-research-data

- Site web :
  - https://gricad.gricad-pages.univ-grenoble-alpes.fr/cellule-data-stewardship/web/

CELLULE DATA
GRENOBLE ALPES

Ressources

- Marie Puren. *Créer son plan de gestion des données.* École thématique. Lille, France. 2021 : https://hal.archives-ouvertes.fr/hal-03183724

- Bonnes pratiques : *Adopter un plan de gestion des données* https://www.datacc.org/bonnes-pratiques/adopter-un-plan-de-gestion-des-donnees/

CELLULE DATA
GRENOBLE ALPES