Guide simplifié

Gestion des données de vos recherches à l'usage des porteurs de projets

Frédéric de Lamotte et al. *v 1* *mai 2025*

Message à destination des porteurs de projets.

Ce document vise à synthétiser les recommandations pour une meilleure gestion des données des projets de recherche. Ceci est ma proposition pour vous faciliter la vie.

Quatre pages pour balayer les différents aspects de la gestion des données suivies d'**une** page de ressources, aucune excuse pour ne pas les lire !

Vous pouvez contribuer à son amélioration en nous faisant des retours.

RECOMMANDATIONS DE BASES POUR LA GESTION DES DONNÉES.

Vous devez aborder les sujets suivants :

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

a. Comment les nouvelles données seront-elles collectées ou produites et/ou comment les données existantes seront-elles réutilisées ?

Expliquez quelles méthodologies ou quels logiciels seront utilisés si de nouvelles données sont collectées ou produites (*comme un matériel et méthode, mais en mieux* ...).

- *Pointez vers les plateformes qui ont produit les données obtenues à l'extérieur (pointez vers le PGD d'entité ou de plateforme s'il existe).*
- *Fournissez les méthodes que vous avez utilisées pour vos propres expériences*
- *Citez (précisément) les logiciels, précisez leurs versions et les formats utilisés, etc*

Indiquez toute contrainte concernant la réutilisation des données existantes.

- *Contraintes juridiques : Respect des règles comme le RGPD, notamment pour les données personnelles.*
- *Propriété intellectuelle : Droits d'auteur ou restrictions imposées par des contrats ou des accords avec des partenaires non académiques*
- *Contraintes éthiques : Respect des principes éthiques, surtout pour les populations vulnérables ou les données collectées dans un cadre spécifique.*
- *Problèmes techniques : ex : Formats obsolètes nécessitant une adaptation avant réutilisation.*

Expliquez comment la provenance des données sera documentée.

- *Origine des données : Identifier clairement si les données sont collectées, produites ou réutilisées, en mentionnant leurs sources (entrepôt d'origine, version ou date d'accès).*
- *Traçabilité : Décrire les étapes de création, modification et analyse des données.*
- *Documentation : Fournir des informations sur la méthodologie utilisée, les outils employés et les transformations appliquées aux données.*
- *Standards : Utiliser des normes reconnues pour garantir une documentation claire et interopérables (ex. PROV pour la biologie).*
    - *Pour cette étape, dessiner un workflow du cheminement des données permet de clarifier la situation et fournit une vue synthétique*

b. Quelles données seront collectées ou produites ?

Donnez des détails sur le type de données :

- *Par exemple, données numériques (bases de données, feuilles de calcul), textuelles (documents), images, audio, vidéo ou multimédias.*

Précisez le format des données, et favorisez les formats ouverts et standards.

- *Par exemple, privilégiez le txt au doc, le csv à l'xls (voir facile.cines.fr pour des exemples précis)*

Indiquez les volumes de données :

- *Exprimez-les en espace de stockage requis en termes de volume et de temps !*

2. Documentation et qualité des données

a. Quelles métadonnées et documentations accompagneront les données ? (par exemple, la méthodologie de collecte des données et le mode d'organisation des données)

Indiquez quelles métadonnées seront fournies pour aider les autres à identifier et à découvrir les données.

- *Précisez les standards de métadonnées utilisés (par exemple DublinCore, DDI, TEI, EML, MIAPPE).*
- *Cette étape peut sembler compliquée au début. Pour vous aider, consultez des articles récents de votre domaine pour repérer les standards utilisés, ou visitez re3data.org pour découvrir les exigences des entrepôts de données.*
- *Pour cette étape particulièrement, faites vous aider par des spécialistes de la gestion des données*

Expliquez comment les données seront organisées pendant le projet

- *Centraliser les fichiers dans un espace partagé sécurisé (ex. : cloud institutionnel, cluster de calcul) pour éviter les duplications et garantir un accès fiable et sécurisé.*
- *Structure des dossiers : répertoires hiérarchiques avec des sous-dossiers spécifiques (ex. : Données brutes, Données traitées, Analyses). Dans chaque sous-dossier, ajoutez un fichier readme qui explique son contenu.*
- *Convention de nommage : mettez vous d'accord sur des noms de fichiers clairs, descriptifs et cohérents (ex. : ProjetX_Analyse_20250131.csv), sans caractères spéciaux ni espaces (_ est votre ami).*
- *Documentation et métadonnées. Associer des métadonnées aux fichiers (ex. : auteur, date de création, description du contenu) pour faciliter leur réutilisation. La façon la plus simple est d'ajouter un autre fichier compagnon dans le dossier (metadata.txt).*

Réfléchissez à la documentation nécessaire pour permettre la réutilisation. Cela peut inclure des informations sur la méthodologie de collecte des données, des détails analytiques, des définitions des variables, des unités de mesure, etc.

- *Là encore, pensez aux fichiers compagnons, ou faites des renvois vers les protocoles qui seraient dans une espace documentaire ou un cahier de labo électronique.*

b. Quelles mesures de contrôle de la qualité des données seront utilisées ?

Vérification de l'exactitude et de la cohérence.

- *S'assurer que les données sont exemptes d'erreurs (documenter la procédure de contrôle qualité utilisée) et cohérentes entre différents ensembles ou champs (par exemple, absence de doublons - vérifiez vos fiches Excel avec OpenRefine).*

Exhaustivité et précision

- *Vérifier que toutes les données requises sont présentes et correctement renseignées*

Standardisation

- *Utiliser des formats standards et homogènes pour garantir la compatibilité et l'interopérabilité (ex. : formats de date uniformes et basés sur la nome ISO, codage standardisé, osons le mot « ontologies ») .*

Validation des méthodes analytiques.

- *Vérifier que les méthodes utilisées pour collecter ou traiter les données répondent aux critères définis (par exemple, robustesse, spécificité) .*

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 ?

- *Réfléchissez dès le début du projet pour réserver (et budgétiser) ces espaces. La question vous est posée dans le PGD (la vie es tbien faite;-) - regardez les ressources citées en 7).*

b. Comment la sécurité des données et la protection des données sensibles seront-elles assurées durant la recherche ?

- *Si vous choisissez bien vos solutions de stockage (ex Nextcloud + Cluster de calcul), ces aspects seront gérés par des spécialistes, c'et toujours ça de soucis en moins !*

c. Comment le code source des logiciels, scripts, modèles sera-t'il géré lors du projet ?

- *Utiliser si possible une forge logicielle institutionnelle qui se trouve dans la fédération d'identité et qui accessible a l'ensemble de l'ESR ! (ex : forge.inrae.fr)*
- *Formez vous a l'utilisation d'une forge, des sessions d'initiation et de formation sont organisés dans les centres INRAE et dans tout l'ESR*

4. Exigences légales et éthiques, codes de conduite

a. Si des données personnelles sont traitées, comment le respect de la législation sur les données personnelles et la sécurité des données sera-t-il garanti ?

- *Obtenez un consentement éclairé pour la conservation et/ou le partage des données personnelles.*
- *Utilisez la procédure simplifiée de déclaration grâce au logiciel https://securite-rgpd.inrae.fr*
- *Assurez-vous de l'anonymisation des données personnelles, car les données anonymisées ne sont plus considérées comme personnelles et échappent au RGPD.*
- *Une alternative possible est la pseudonymisation, qui remplace les informations identifiantes par des pseudonymes (processus réversible), tout en restant conforme au RGPD.*
- *Précisez s'il existe une procédure d'accès contrôlé pour les utilisateurs autorisés à consulter les données personnelles.*

b. Comment les autres questions juridiques, telles que les droits de propriété intellectuelle et la propriété, seront-elles gérées ? Quelle législation est applicable ?

- *Contactez les spécialistes de la valorisation et du partenariat dès le montage du projet pour gérer ces aspects. Ils ont parfois un langage bizarre mais ceux sont vos amis (voire ressources ci-après).*

5. Partage des données et préservation à long terme

a. Comment et quand les données seront-elles partagées ? Existe-t-il d'éventuelles restrictions ou raisons d'embargo pour le partage des données ?

- *Les données financées par des fonds publiques sont … publiques. Reste à définir quand et comment elles seront partagées. Vous devez privilégier les entrepôts thématiques de confiance. À défaut de solution, considérez Recherche Data Gouv. Ces entrepôts fournissent un accès 24/7 à vos données et fournissent des identifiants uniques pérennes. Cela favorise le partage et la découverte de vos données.*
- *Il est possible de prévoir une durée d'embargo : délais entre dépôt et ouverture au public (choisir une durée réaliste et honnête : moins de 24 mois).*
- *Si vous utilisez une forge logicielle rendez le dépôt public, il sera archivé automatiquement par « Software Heritage » et bénéficiera d'un SWID (elle est pas belle la vie ?)*

b. Comment les données destinées à la préservation seront-elles sélectionnées, et où seront-elles conservées à long terme (par exemple, dans un dépôt ou une archive) ?

- *Le partage des données brutes est obligatoire à la fin du projet (volet éthique du plan national Science Ouverte). Il faut aussi partager les données ayant un fort potentiel de réutilisation, y compris par des communautés distantes (ex : les données de dates de fleurissement des pommiers intéressent les climatologues).*

c. Quels outils ou logiciels seront nécessaires pour accéder aux données et les utiliser ?

- *N'oubliez pas de partager vos données sous des formats standards et ouverts ne nécessitant pas de logiciel propriétaires. Les entrepôts nationaux et internationaux préconisent en général un format de stockage adapté au type de vos données.*

6. Responsabilités en matière de gestion des données et ressources

a. Qui (par exemple rôle, poste, institution) sera responsable de la gestion des données ?

- *Il faut attribuer et distribuer les responsabilités pour être sur que tous les aspects du cycle de vie des données sont bien sous la responsabilité d'une ou plusieurs personnes.* ***Le PGD est un très bon outil pour s'en assurer !***

b. Quelles ressources (par exemple financières ou en termes de temps) seront consacrées à la gestion des données ?

- *Il faut prévoir les ressources matérielles pour toute la durée du projet (tant que toutes les données n'auront pas été déposées dans des entrepôts thématiques).*
- *L'aspect ressources humaines est aussi à intégrer (monter en compétence en interne ou recrutement).*

7. Ressources

Certains des éléments de cette liste sont spécifiques à INRAE (accès après identification)
Pleins de liens cliquables à explorer !

Accompagnement calcul et stockage à INRAE (accompagnement-numerique@inrae.fr)
Demande de ressources INRAE (https://ariane.inrae.fr/block?id=ariane_sc_category&catalog_id=-1)
    Aide à la curation des données (datainrae@inrae.fr)
    Aide au PGD (pgd@inrae.fr)
Formations en ligne
    La formation OSCAR d'INRAE est maintenant accessible à tous
    L'Institut Français de Bioinformatique vous accompagne également

Décrire ses données de recherche
Choisir un espace de travail collaboratifs pendant le projet
Comprendre les règles de nommage ou aussi
Choisir des formats ouverts pour vos fichiers
Choisir des standards de métadonnées
    Comprendre les métadonnées
    Deux répertoires de tous les standards : FairSharing et DCC
    Approche complémentaire : trouver un entrepôt dans votre domaine et suivre ses recommandations
Choisir la bonne licence
    La politique française en matière de licence
    Choisir parmi les licences « creative commons » pour vos données
    Choisir des licences pour vos codes
Le RGPD facile grâce à la CNIL

Liste des 300 plateformes de production de données de l'ESR français
Liste des entrepôts de confiance où partager vous données
    Liste exhaustive des entrepôts disponibles : https://www.re3data.org/

Organisation des fichiers pendant le projet
https://gricad.gricad-pages.univ-grenoble-alpes.fr/cellule-data-stewardship/web/organiser/bonnes_pratiques/?t
https://doranum.fr/stockage-archivage/comment-nommer-fichiers_10_13143_wgqw-aa59/

Pour allez plus loin sur tous ces sujets
La gestion des données et aussi ici (en français)
Et en anglais: RDMKit
Un site web qui balaye tous ces points (idéal pour un doctorant qui débute)

8. Contributeurs

Frédéric de Lamotte - 0000-0003-4234-1172
contact : frederic.de-lamotte@inrae.fr

Véronique Stoll - 0000-0002-4118-3574
Hélène Chiapello - 0000-0001-5102-0632
Julien Cufi - 0000-0001-9149-5413