



## INSA de Toulouse: ANR - Modèle de PGD structuré (français)

### Description des données et collecte ou réutilisation de données existantes

Description générale du produit de recherche

#### *Recommandations:*

- Donnez des détails sur le type de données : par exemple numérique (bases de données, tableurs), textuel (documents), image, audio, vidéo, et/ou médias composites.
- Expliquez comment les données pourraient être réutilisées dans d'autres contextes. Les identifiants persistants (PID) devraient être appliqués de manière à ce que les données puissent être localisées et référencées de façon fiable et efficace. Ces PID aident aussi à comptabiliser les citations et les réutilisations.
- Indiquez s'il sera envisagé d'attribuer aux données un PID. Typiquement, un entrepôt pérenne digne de confiance attribuera des identifiants persistants.

#### *Recommandations:*

Dans cette section, on vous demande de décrire le type de données qui seront générées ou réutilisées dans le cadre de votre projet : par exemple numérique (bases de données, tableurs), textuel (documents), image, audio, vidéo, et/ou médias composites... Le format demandé est celui sous lequel les données seront stockées et mises à disposition : PDF, xls, doc, txt, rdf... Nous vous recommandons de privilégier des formats ouverts (non propriétaires) et pérennes. Pour vérifier que les formats utilisés sont pérennes et archivables, vous pouvez utiliser l'outil du CINES : Enfin, la volumétrie de vos données peut être exprimées en espace de stockage requis (octets), et/ou en quantité d'objets, de fichiers, de lignes, de colonnes... Il peut être assez difficile de donner la volumétrie des données en début de projet. Vous pouvez donc seulement fournir une estimation et faire évoluer cette partie à chaque nouvelle version du PGD.

Est-ce que des données existantes seront réutilisées ?

#### *Recommandations:*

- Énoncer les éventuelles contraintes à la réutilisation des données préexistantes.
- Indiquer brièvement le cas échéant, les raisons pour lesquelles l'utilisation de sources de données existantes a été envisagée mais écartée.

Comment seront produites/collectées les nouvelles données ?

#### *Recommandations:*

- Expliquer quelles méthodologies ou quels logiciels seront utilisés si de nouvelles données sont recueillies ou produites.
- Expliquer comment la provenance des données sera documentée.

### Documentation et qualité des données

Quelles métadonnées et quelle documentation (par exemple mode d'organisation des données) accompagneront les données ?

#### *Recommandations:*

- Indiquez quelles métadonnées seront fournies pour aider à la recherche et à l'identification des données.
- Indiquez quelles standards de métadonnées seront utilisés (par exemple DDI, TEI, EML, MARC, CMDI).
- Utiliser les standards de métadonnées de communauté lorsque ceux-ci existent.
- Indiquez comment les données seront organisées au cours du projet, en mentionnant par exemple conventions de nommage, contrôle de version et structures des dossiers. Des données bien classées et gérées de façon cohérente seront plus faciles à retrouver, à comprendre et à réutiliser.
- Pensez à la documentation qui serait nécessaire pour permettre une réutilisation des données. Il peut s'agir notamment de l'information sur la méthodologie utilisée pour collecter les données, sur les procédures et méthodes d'analyse utilisées, sur la définition des variables, des unités de mesure, etc.
- Tenez compte de la façon dont ces informations seront obtenues et enregistrées par exemple dans une base de données avec des liens vers chacun des fichiers, dans un fichier texte de type « lisez-moi », dans les en-têtes de fichiers, dans un livre de référence (« code book ») ou dans les cahiers de laboratoire.

#### *Recommandations:*

L'objectif de cette section est d'assurer une bonne compréhension des données afin de permettre une réutilisation ultérieure.

Si vous utilisez des standards de métadonnées (par exemple DDI (Data Documentation Initiative), TEI, EML (Ecological Metadata Language), MARC, CMDI, etc. ; le site [Disciplinary Metadata](#) donne les métadonnées par disciplines et par types de ressources) pour décrire les données brutes comme les données délivrables vous pouvez l'indiquer ici. Si vous ne pouvez pas utiliser de standards reconnus, il faut préciser les métadonnées que vous utilisez. L'outil [Datacite Metadata Generator](#) vous permet de générer un fichier de métadonnées.

Pour accompagner vos jeux de données, il vous sera demandé de produire un fichier Readme, qui servira à décrire les fichiers, leurs conventions de nommage, le contrôle des versions, la structure des dossiers, éventuellement les logiciels à utiliser pour les consulter, etc. Vous pouvez également accompagner vos données de l'ensemble la documentation associée : questionnaire d'enquête, dictionnaire des variables, protocole, ...

L'objectif est que chaque jeu de données soit accompagné de toutes les informations nécessaires pour que les données puissent être accessibles, interprétables et réutilisables par une personne n'ayant pas participé à l'étude qui les a produites.

Pour vous aider, le guide "Décrire ses jeux de données dans les règles du FAIR" : <https://doi.org/10.5281/zenodo.8376581>

Quelles seront les méthodes utilisées pour assurer la qualité scientifique des données ?

*Recommandations:*

- Expliquer comment la qualité et la conformité de la collecte des données seront contrôlées et documentées. Il s'agit là de préciser les processus comme la calibration, la répétition des échantillons ou des mesures, la capture standardisée des données, la validation de saisie des données, la revue par les pairs, ou la représentation basée sur des vocabulaires contrôlés.

*Recommandations:*

Pour décrire la démarche de contrôle qualité que vous allez adopter vous pouvez vous appuyer sur l'article « [Qualité des données](https://doi-org.gorgone.univ-toulouse.fr/10.51257/a-v2-h3700) » sur la base Techniques de l'Ingénieur (<https://doi-org.gorgone.univ-toulouse.fr/10.51257/a-v2-h3700>) ou sur le guide « [Traçabilité des activités de recherche](https://qualite-en-recherche.cnrs.fr/gt/tracabilite-des-activites-de-recherche/) » du CNRS (<https://qualite-en-recherche.cnrs.fr/gt/tracabilite-des-activites-de-recherche/>)

## Exigences légales et éthiques, code de conduite

Quelles seront les mesures appliquées pour assurer la protection des données à caractère personnel ?

*Recommandations:*

- Lorsque vous manipulez des données à caractère personnel, veillez à ce que les lois sur la protection des données (par exemple, RGPD) soient appliquées, notamment :
  - o Obtenir un consentement éclairé pour la préservation et/ou le partage de données personnelles.
  - o Envisager l'anonymisation des données personnelles pour la préservation et/ou le partage (des données correctement anonymisées ne sont plus considérées comme des données personnelles).
  - o Envisager la pseudonymisation des données personnelles (la principale différence avec l'anonymisation est que la pseudonymisation est réversible).
  - o Envisagez le chiffrement des données, qui est considéré comme un cas particulier de pseudonymisation (la clé de cryptage doit alors être stockée séparément des données, par exemple par un tiers de confiance).
  - o Expliquez si une procédure d'accès spécifique a été mise en place pour les utilisateurs autorisés à accéder aux données personnelles.

*Recommandations:*

Si votre étude comporte plus spécifiquement des données à caractère personnel il vous faudra entrer en contact avec le DPO de l'INSA (contact-dpo@insa-toulouse.fr) pour qu'il certifie la conformité du projet avec le RGPD, notamment en ce qui concerne le consentement des participants pour la conservation et le partage de leurs données après anonymisation. Il vous sera également demandé d'expliquer de quelle façon les données seront anonymisées et éventuellement cryptées pendant le stockage.

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 ?

*Recommandations:*

- Expliquez qui sera le propriétaire des données, c'est-à-dire qui aura le droit d'en contrôler l'accès :
  - o Expliquez quelles conditions d'accès s'appliqueront aux données ? Les données seront-elles librement accessibles, ou des restrictions seront-elles appliquées ? Si oui, lesquelles ? Envisager l'utilisation de licences concernant l'accès et la réutilisation des données.
  - o Assurez-vous de couvrir, dans l'accord de consortium, ces questions de droits de contrôle d'accès aux données pour les projets multipartenaires et en cas de propriété partagée des données.
- Indiquez si les droits de propriété intellectuelle (par exemple la directive bases de données, droits sui generis) sont affectés. Dans l'affirmative, veuillez expliquer lesquels et comment cela sera-t-il traité.
- Indiquez s'il y a des restrictions sur la réutilisation des données fournies par des tiers.

*Recommandations:*

Pour les données générées au cours du projet, le propriétaire est en principe précisé dans l'accord de consortium. Vous pouvez donc reprendre les éléments définis dans le document en précisant les conditions d'accès qui s'appliqueront aux données : seront-elles librement accessibles, ou des restrictions seront-elles appliquées ? Si oui, lesquelles ? Pour permettre l'accès et la réutilisation des données quelles licences leurs seront appliquées ?

Exemple : *Pour les données produites dans le cadre du projet de recherche, les partenaires de recherche sont conjointement propriétaires des données tel que mentionné dans l'accord de consortium. A la fin du projet, les données qui seront en libre accès seront sous licence X.*

Pour les données préexistantes et que vous réutilisez dans le projet de recherche il faut en préciser la source des données ainsi que les droits de propriété intellectuelle (données libres de droits ? sous licence ?)

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

*Recommandations:*

- Déterminer si les questions d'éthique ont une incidence sur la façon dont les données sont stockées et transférées, qui peut les voir ou les utiliser, et quelle durée de conservation leur seront-elles appliquées. Démontrer que ces aspects sont bien pris en compte et planifiés.
- Adoptez les codes de conduite nationaux et internationaux et le code d'éthique institutionnel et vérifiez si une revue des pratiques (par exemple par un comité d'éthique) est requise pour ce qui concerne la collecte de données dans le cadre du projet de recherche.

*Recommandations:*

Si les données recueillies font l'objet de questions d'éthique, elles devront subir un traitement particulier en conséquent. Si l'approbation d'un comité d'éthique est nécessaire vous pouvez l'indiquer ici avec ses recommandations. Pour ces questions vous pouvez contacter la référente intégrité scientifique de l'établissement.

## Traitement et analyse des données

Comment et avec quels moyens seront traitées les données ?

## Stockage et sauvegarde des données pendant le processus de recherche

Comment les données seront-elles stockées et sauvegardées tout au long du projet ?

*Recommandations:*

- Décrire l'endroit où les données seront stockées et sauvegardées au cours du processus de recherche et la fréquence à laquelle la sauvegarde sera effectuée. Il est recommandé de stocker les données dans au moins deux lieux distincts.
- Privilégier l'utilisation de systèmes de stockage robustes, avec sauvegarde automatique, tels que ceux fournis par les services informatiques de l'institution d'origine. Le stockage des données sur des ordinateurs portables, des disques durs externes, ou des périphériques de stockage tels que des clés USB n'est pas recommandé.
- Expliquez comment les données seront récupérées en cas d'incident.
- Expliquer qui aura accès aux données au cours du processus de recherche et comment l'accès aux données est contrôlé, en particulier dans le cadre de recherches menées en collaboration.
- Tenez compte de la protection des données, en particulier si vos données sont sensibles (par exemple données à caractère personnel, politiquement sensibles des informations ou secrets commerciaux). Décrire les principaux risques et la façon dont ils seront gérés.
- Expliquer quelle politique institutionnelle de protection des données institutionnelles est mis en œuvre.

*Recommandations:*

Il vous est demandé ici de décrire l'endroit où les données seront stockées et sauvegardées au cours du projet et la fréquence à laquelle la sauvegarde sera effectuée (ici il est question du stockage des données chaudes pendant la durée du projet de recherche. Le stockage après la fin du projet sera abordé à la question 5). Afin d'éviter tout risque de perte ou de faille de sécurité il est recommandé de stocker les données dans au moins deux lieux distincts.

Il vous est recommandé de privilégier l'utilisation de systèmes de stockage robustes, avec sauvegarde automatique (outils institutionnels par ex.). Il n'est pas recommandé de stocker vos données sur des ordinateurs portables, des disques durs externes, ou des périphériques de stockage tels que des clés USB.

## Partage des données et conservation à long terme

Comment les données seront-elles partagées ?

*Recommandations:*

- Détailler le format des données : la manière selon laquelle les données sont codées pour le stockage, généralement reflétée par l'extension du nom de fichier (par exemple PDF, XLS, DOC, TXT, ou RDF).
- Justifier l'utilisation de certains formats. Par exemple, les choix d'un format peuvent être guidés par l'expertise du personnel de l'organisme, ou par une préférence pour les formats ouverts, par les standards de format acceptés par les entrepôts de données, par l'usage largement répandu dans une communauté de recherche ou par le logiciel ou l'équipement qui sera utilisé.
- Privilégier les formats standards et ouverts car ils facilitent le partage et la réutilisation à long terme des données (plusieurs catalogues fournissent des listes de ces « formats préférés »).
- Donnez des détails sur les volumes (qui peuvent être exprimés en espace de stockage requis (octets), et/ou en quantités d'objets, de fichiers, de lignes, et colonnes).
- Expliquez comment les données pourront être retrouvées et partagées (par exemple, par dépôt dans un entrepôt de données de confiance, indexé dans un catalogue, par l'utilisation d'un service de données sécurisé, par traitement direct des demandes de données, ou l'utilisation de tout autre mécanisme).
- Expliquez à quel moment les données seront rendues disponibles. Indiquer les délais de publication prévus. Expliquez si une utilisation exclusive des données est revendiquée et, dans l'affirmative, pour quelle raison et pour combien de temps. Indiquez si le partage des données sera différé ou limité, par exemple pour des raisons de publication, pour protéger la propriété intellectuelle, ou dépôt de brevets.
- Indiquez qui pourra utiliser les données. S'il s'avère nécessaire de restreindre l'accès pour certaines communautés ou d'imposer un accord pour le partage de données, expliquez comment et pourquoi. Expliquez les mesures qui seront prises pour dépasser ou minimiser ces restrictions.
- Décrire les utilisations (et/ou les utilisateurs) prévisibles des données dans un cadre de recherche.
- Indiquez où les données seront déposées. Si aucun entrepôt reconnu n'est proposé, démontrer dans le DMP que les données pourront être prises en charge efficacement au-delà de la durée de financement du projet. Il est recommandé de démontrer que les politiques des entrepôts et les procédures de dépôts (y compris les normes de métadonnées, et coûts mis en œuvre) ont été vérifiés.

- Indiquez si les utilisateurs potentiels ont besoin d'outils spécifiques pour l'accès et la (ré)utilisation des données. Tenez compte de la durée de vie des logiciels nécessaires pour accéder aux données.
- Indiquez si les données seront partagées via un entrepôt, si les demandes d'accès seront traitées en direct, ou si un autre mécanisme sera utilisé ?

*Recommandations:*

Dans la première version du PGD vous pouvez renseigner ce champ a minima : indiquer que vous prenez en compte ces questions pour répondre aux exigences de l'agence de financement et que vous en préciserez les modalités dans les versions suivantes du PGD.

Pour la diffusion des données après la fin du projet les agences de financements et l'INSA recommandent de passer par un entrepôt de données, de préférence labellisé entrepôt de confiance. Vous n'avez pas besoin de donner ici le nom de l'entrepôt choisi, seulement de préciser la solution de stockage pérenne choisie. Si vous faites un autre choix que l'entrepôt de données, il vous sera demandé de prouver que les données pourront être prises en charge efficacement au-delà de la durée du financement du projet. Si vos données feront l'objet d'un embargo avant diffusion il faut l'indiquer, en en précisant la durée et la raison (données sensibles ou personnelles, obligations contractuelles, etc.).

Dans le cadre du partage de vos données et de leur réutilisation le ministère préconise l'emploi de la « Licence Ouverte / Open License » d'Etalab.

**Attention**, depuis la Loi pour une République Numérique (Art. L312-1-1 CRPA), les données achevées produites au sein d'établissements publics doivent obligatoirement être publiées en ligne, selon le principe "aussi ouvert que possible, aussi fermé que nécessaire".

Comment les données seront-elles conservées à long terme ?

*Recommandations:*

- Définissez le plan de préservation des données et fournir l'information sur la durée d'archivage pérenne des données.
- Indiquez quelles données ne doivent pas être divulguées ou doivent être détruites pour des raisons contractuelles, légales, ou réglementaires.
- Indiquez comment il sera décidé quelles données garder. Décrire les données qui seront à préserver à long terme.

*Recommandations:*

A la fin du projet toutes les données n'auront pas vocation à être conservées et diffusées, quelles données ne doivent pas être divulguées ou seront détruites (pour des raisons contractuelles, légales, réglementaires, etc.) ? Sur quels critères déciderez-vous quelles données vous conserverez et combien de temps ? Attention, la durée de conservation de certaines données peut être contrainte.

C'est dans cette section que vous pouvez préciser où les données seront déposées. Si aucun entrepôt de confiance (liste sur le site de CoreTrustSeal ou par le répertoire Re3data) ou entrepôt disciplinaire ne correspond à votre projet vous pouvez déposer vos données sur RechercheDataGouv, l'entrepôt national créé par le MESR pour répondre aux besoins des institutions françaises.