

Modèle de plan de gestion des données : Les flux de travail en science ouverte

Résumé

Ce modèle est conçu comme un guide de rédaction pour les plans de gestion des données (PGD) comportant des flux de travail en science/érudition ouverte, employant des méthodologies mixtes en sciences sociales et produisant des jeux de données quantitatives et qualitatives. Les questions mettent l'accent sur le partage et la réutilisation des données tout au long du projet ainsi qu'au stade final de la publication. Ce modèle de PGD sera très utile aux chercheurs qui travaillent dans le cadre d'un partenariat entre multiples établissements et qui ont déjà rempli une demande de financement et un certificat d'éthique de la recherche. Le PGD est un document évolutif : n'oubliez pas de revoir votre PGD tout au long du projet de recherche pour mettre à jour ou réviser vos réponses.

Les questions ne sont pas forcément pertinentes pour tous les projets de recherche. Nous vous encourageons à répondre au plus grand nombre de questions possible, mais c'est à vous et à votre équipe de décider quelles questions et réponses s'appliquent à votre flux de travail.

Ce modèle est également disponible sur l'espace communautaire Zenodo de Meaningful Data Counts : <https://doi.org/10.5281/zenodo.4092122>.

Détails administratifs

Auteur (s) du modèle : Erica Morissette, Lina Harper, Université d'Ottawa; Isabella Peters, ZBW Leibniz Information Centre for Economics, CAU Kiel University; Felicity Tayler, Stefanie Haustein, Université d'Ottawa

Publié : 19 avril 2021

DOI : [10.5281/zenodo.4701045](https://doi.org/10.5281/zenodo.4701045)

Contact : Le Réseau Portage - portage@engagedri.ca, reseauportage.ca

Licence : [Attribution - Pas d'Utilisation Commerciale 4.0 International \(CC BY-NC 4.0\)](https://creativecommons.org/licenses/by-nc/4.0/)



Version :

Version	Date	Changements
1.0	2021-04-09	Préparé pour la publication inaugurale.
1.1	2021-04-19	Attribution ajustée.

Responsabilités et ressources

Qui sera responsable de la gestion des données? Le chercheur principal sera-t-il entièrement responsable de celle-ci pendant et après le projet, ou celle-ci sera-t-elle divisée entre les membres de l'équipe ou d'un partenaire?

Attribution des responsabilités : Une fois terminé, votre plan de gestion des données décrira les activités importantes de votre projet. Désignez les personnes ou les organisations qui seront responsables de la réalisation de ces parties de votre plan de gestion des données. Cette étape peut également inclure le calendrier associé à ces responsabilités du personnel et toute formation nécessaire pour préparer le personnel à ces tâches.

Qui remplacera le chercheur principal si celui-ci quitte le projet? Qui assumera ses responsabilités temporairement jusqu'à l'arrivée d'un nouveau chercheur principal?

Planification de la relève : Le chercheur principal est généralement chargé de maintenir les normes d'accessibilité des données pour l'équipe. Réfléchissez à la personne qui répondra aux questions concernant l'accès à l'information ou l'autorisation d'accès aux données dans le cas où le chercheur principal quitterait le projet. Habituellement, le cochercheur principal prend en charge ces responsabilités.

Précisez une stratégie de relève pour ces données si des personnes responsables des données quittent le projet (par exemple, un étudiant de cycle supérieur qui quitte le projet après avoir obtenu son diplôme). Décrivez le processus à suivre dans le cas où le chercheur principal quitte le projet. Dans certains cas, un cochercheur ou le département ou la division qui supervise cette recherche en assumera la responsabilité.

Énumérez toutes les ressources prévues pour la gestion des données nécessaires à la réalisation de votre projet. Quels matériels, quels logiciels et quelles ressources humaines vous faut-il? Quel est votre budget prévisionnel?

Budgétisation : Les achats les plus courants sont les disques durs, le stockage informatique ou l'accès aux logiciels. L'outil de calcul des coûts de gestion des données de l'Université de technologie de Delft est utile pour déterminer la part de travail humain (ÉTP) qui doit être allouée à la gestion des données de recherche (*lien en anglais*).

Collecte de données

Quels types de données recueillerez-vous, créez-vous, couplerez-vous, acquérez-vous ou consignerez-vous?

Types de données : Vos données de recherche peuvent comprendre des ressources numériques, des codes logiciels, des fichiers audio, des fichiers images, des données vidéo, numériques, textuelles, tabulaires, des données de modélisation, des données spatiales et des données d'instrumentation.

Répondez aux questions suivantes concernant les formats de fichiers :

- A. Sous quels formats de fichier collecterez-vous les données (p. ex. DOC, CSV, JPG, MOV)?**
- B. Ces formats de fichiers sont-ils faciles à partager avec d'autres chercheurs de différentes disciplines?**
- C. Si l'un des formats choisis devenait désuet (ou n'était plus pris en charge), comment garantiriez-vous l'accès aux données de recherche?**
- D. Vos données doivent-elles être copiées sur un nouveau média ou une plateforme d'infonuagique ou converties dans un format de fichier différent lorsque vous stockez ou publiez vos jeux de données?**

Les formats de fichiers propriétaires nécessitant l'utilisation de logiciels ou de matériel spécialisés ne sont pas recommandés, mais peuvent être nécessaires pour certaines méthodes de collecte ou d'analyse des données. L'utilisation de formats de fichiers ouverts ou de formats standard (par exemple, ceux largement utilisés par une communauté donnée) est préférable dans la mesure du possible. Pour en savoir plus sur les formats de fichiers, consultez les sites suivants : [Bibliothèque et Archives Canada](#), la [Bibliothèque de l'Université de la C.-B.](#) (*lien en anglais*) ou le [UK Data Service](#) (*lien en anglais*).

Répondez aux questions suivantes concernant les conventions de dénomination :

- A. Comment allez-vous structurer, nommer et gérer les versions de vos fichiers pour aider une personne en dehors de votre équipe de recherche à comprendre comment vos données sont organisées?**
- B. Écrivez étape par étape votre flux de travail idéal pour le partage de fichiers entre les membres de l'équipe de recherche.**
- C. Quels outils ou stratégies utiliserez-vous pour documenter votre flux de travail au fur et à mesure de son évolution au cours du projet?**

Nomenclature et version de fichier : Il est important de garder une trace des différentes copies ou versions de fichiers, des fichiers sauvegardés en différents formats ou à des endroits différents, et des informations croisées entre les fichiers. Ce processus s'appelle la « gestion de versions ». Les structures de fichiers logiques, les conventions de nomenclature informatives et les indications claires sur les versions des fichiers contribuent à une meilleure utilisation de vos données pendant et après votre projet de recherche. Ces pratiques permettront de s'assurer que vous et votre équipe de recherche utilisez la version appropriée de vos données, et de minimiser la confusion concernant les copies sur différents ordinateurs ou sur différents supports. Pour en savoir plus sur la nomenclature et la gestion des versions de fichier, consultez le site de la [Bibliothèque de l'Université de la C.-B.](#) ou celui du [UK Data Service](#) (*liens en anglais*).

N'oubliez pas que vous pouvez adapter ce flux de travail et mettre le PGD à jour tout au long du projet.

L'utilisation d'une [feuille de travail pour la convention de nomenclature](#) peut être très utile (*lien en anglais*). Assurez-vous que la convention n'utilise que des caractères alphanumériques, des tirets et des traits de soulignement. En général, les noms de fichiers doivent comporter 32 caractères ou moins et contenir la date et le numéro de version (par exemple : "P1-MUS023_2020-02-29_051_raw.tif" and "P1-MUS023_2020-02-29_051_clean_v1.tif")

Flux de travail pour les documents : Avez-vous réfléchi à la manière dont vous allez saisir, sauvegarder et partager avec votre équipe les étapes de votre travail et de vos projets? Vous pouvez créer un document d'intégration pour vous assurer que tous les membres de l'équipe adoptent les mêmes flux de travail ou utiliser des outils de gestion des flux de travail comme [OSF](#) ou [GitHub](#) (*liens en anglais*).

Documentation et métadonnées

De quel matériel de soutien et de quelle documentation (par exemple, un fichier LISEZMOI ou README) les membres de votre équipe et les futurs chercheurs auront-ils besoin pour naviguer et réutiliser vos données sans ambiguïté?

Documentation des données : Il est fortement recommandé d'inclure un fichier LISEZMOI ou README avec tous les jeux de données (ou éléments semblables) pour aider à comprendre les méthodes de collecte et de traitement des données, les conventions de nomenclature et la structure des fichiers.

En règle générale, une bonne documentation des données comprend des informations sur l'étude, des descriptions des données et toute autre information contextuelle nécessaire pour rendre les données utilisables par d'autres chercheurs. Voici d'autres éléments que vous devez documenter, le cas échéant : la méthodologie de recherche utilisée, les définitions des variables, les vocabulaires, les systèmes de classification, les unités de mesure, les hypothèses formulées, le format et le type de fichier des données, une description des méthodes de saisie et de collecte des données, une explication du codage et de l'analyse des données (y compris les fichiers de syntaxe). Vous pouvez consulter le [Guide rapide de création d'un fichier README pour vos jeux de données](#) de l'Université de la C.-B. pour plus d'informations.

Comment allez-vous documenter la collecte, le traitement et l'analyse des données, dans le cadre de votre flux de travail, afin de créer un matériel d'aide cohérent? Qui sera responsable de cette tâche?

Attribution des responsabilités quant à la documentation : Les rôles individuels et les flux de travail devraient inclure la collecte de la documentation des données comme élément clé.

Il est utile d'établir les responsabilités en matière de gestion des données et de documentation dès le début projet, idéalement avant la collecte et l'analyse des données. Certains chercheurs utilisent la [taxonomie de CRediT](#) pour déterminer les rôles des auteurs au début de chaque projet (*lien en anglais*). Ils peuvent également être utilisés pour rendre les membres de l'équipe responsables de la gestion appropriée des données et de la documentation.

Songez à la manière dont vous allez saisir ces informations et à l'endroit où elles seront enregistrées, afin de garantir l'exactitude, la cohérence et l'intégralité de la documentation.

Envisagez-vous d'utiliser une norme de métadonnées? Quel schéma spécifique pourriez-vous utiliser?

Métadonnées pour les jeux de données : DataCite a développé un schéma de métadonnées pour les jeux de données ouverts (*lien en anglais*). Il énumère un ensemble de champs de métadonnées de base et d'instructions pour rendre les jeux de données facilement identifiables et citables.

Il existe de nombreuses autres normes de métadonnées générales et spécifiques à un domaine. La documentation sur les jeux de données doit être fournie dans l'un de ces formats standard, lisibles par machine et librement accessibles, afin de permettre un échange efficace d'informations entre les utilisateurs et les systèmes. Ces normes sont souvent basées sur des formats de données indépendants du langage, tels que XML, RDF et JSON. De nombreuses normes de métadonnées sont fondées sur ces formats, y compris des normes spécifiques à une discipline. Pour en savoir plus sur les normes de métadonnées, consultez la ressource sur les métadonnées par disciplines du Digital Curation Centre du Royaume-Uni (*lien en anglais*).

Comment ferez-vous en sorte que (a) vos méthodes de collecte de données primaires soient documentées avec transparence et que (b) vos sources de données secondaires (c'est-à-dire les données que vous n'avez pas collectées vous-même) soient facilement identifiées et citées?

Documentez votre processus : Il est utile de consulter régulièrement les membres de l'équipe de recherche afin de saisir les changements potentiels dans la collecte et le traitement des données qui doivent être reflétés dans la documentation.

Stockage et sauvegarde

Énumérez vos besoins anticipés en matière de stockage (par exemple, disques durs, stockage infonuagique, disques partagés). Indiquez la durée d'utilisation prévue de chaque type et les capacités de stockage dont vous pourriez avoir besoin.

Estimation des besoins de stockage de données : Les estimations de l'espace de stockage doivent tenir compte des besoins en matière de versions de fichier, de sauvegardes et de croissance dans le temps.

Si vous collectez des données sur une longue période (par exemple plusieurs mois ou années), votre stratégie de stockage et de sauvegarde des données doit tenir compte de la croissance des données. Précisez vos supports de stockage de sauvegarde dans votre estimation.

Quel calendrier de sauvegarde et de stockage envisagez-vous? À quelle fréquence, dans quels formats et à quel endroit sauvegarderez-vous vos données?

Suivez la règle 3-2-1 pour éviter les pertes de données : Il est important d'avoir un calendrier de sauvegarde régulier — et de documenter ce processus — afin de pouvoir examiner toute modification des données à tout moment au cours du projet. Vous pouvez réduire sensiblement le risque de perdre des données à la suite d'une erreur humaine, d'une catastrophe naturelle ou d'un autre incident en suivant la règle de sauvegarde 3-2-1 : ayez au moins trois copies de vos données; stockez les copies sur deux supports différents; conservez une copie de sauvegarde hors site. Pour en savoir plus sur les pratiques de stockage et de sauvegarde, consultez le site de la [bibliothèque de l'Université de Sheffield](#) et celui du [UK Data Service](#) (*liens en anglais*).

En fonction des exigences du Comité d'éthique de la recherche, quel est le calendrier de conservation prévu pour chaque type de données (brutes, traitées, propres, définitives) au sein de votre équipe? Allez-vous également stocker du code logiciel ou des métadonnées?

Demandez de l'aide : Votre établissement devrait être en mesure de fournir des conseils sur les solutions de stockage local. Demandez de l'aide à votre bibliothèque ou à votre département d'informatique de recherche avancée par rapport à la gestion des données de recherche (GDR).

Les services de partage de fichiers commerciaux indépendants (tels que Google Drive et Dropbox) facilitent l'échange de fichiers, mais ils ne sont pas nécessairement permanents ou sécurisés, et les serveurs sont souvent situés à l'extérieur du Canada. Ces moyens de stockage peuvent contrevenir aux exigences du Comité d'éthique de la recherche ou à d'autres politiques de l'établissement.

La solution idéale est celle qui facilite la coopération et assure la sécurité des données, tout en pouvant être adoptée par les utilisateurs avec un minimum de formation. La transmission des données entre les sites ou au sein des équipes de recherche peut être difficile pour les infrastructures de gestion des données. Le courrier électronique n'est pas une solution robuste ou sûre pour le transfert des données.

- Les **données brutes** sont obtenues directement à partir de l'instrument, de la simulation ou de l'enquête.
- Les **données traitées** résultent d'une certaine manipulation des données brutes afin d'éliminer les erreurs ou les valeurs aberrantes, de préparer les données pour l'analyse, de dériver de nouvelles variables ou d'anonymiser les participants.
- Les **données analysées** sont les résultats d'une analyse qualitative, statistique ou mathématique des données traitées. Elles peuvent être présentées sous forme de graphiques, de diagrammes ou de tableaux statistiques.
- Les **données finales** sont des données traitées qui ont été converties dans un format facile à préserver si nécessaire.

Partage, réutilisation et préservation

Comment vos données (tant brutes que nettoyées) seront-elles rendues accessibles au-delà de la portée du projet et par des chercheurs en dehors de votre équipe?

Aidez les autres à utiliser et citer vos données : Saviez-vous qu'un jeu de données est une production scientifique que vous pouvez mentionner sur votre CV tout comme un article de revue?

Si vous publiez vos données dans un dépôt de données (par exemple, Zenodo, Dataverse, Dryad), elles peuvent être trouvées et réutilisées par autrui. De nombreux dépôts peuvent émettre des identifiants d'objets numériques (DOI) uniques qui facilitent l'identification et la citation des jeux de données.

re3data.org est un répertoire de dépôts de données ouverts potentiels (*lien en anglais*). Consultez vos collègues pour déterminer quels sont les dépôts courants pour le partage de données dans votre domaine. Vous pouvez également vous renseigner sur le soutien à la GDR de votre établissement, soit à la bibliothèque soit au département d'informatique de recherche avancée. En l'absence de soutien local, consultez le [guide sur les options de dépôt de Portage](#).

La préservation numérique est-elle une composante de votre projet et devez-vous planifier l'archivage et la préservation à long terme?

Combien de temps devez-vous conserver vos données? La durée pendant laquelle vous conserverez ou partagerez vos données au-delà de la phase active de votre recherche peut être déterminée par des politiques externes (par exemple, les organismes de financement, les éditeurs de revues), ou par une compréhension de la valeur durable (historique) d'un jeu de données en particulier. La nécessité de publier les données à court terme (c'est-à-dire à des fins de révision par les pairs), d'y accéder à plus long terme pour les réutiliser (pour se conformer aux exigences de financement) ou de les préserver par une conversion et une migration continues des fichiers (pour les données ayant une valeur historique durable) influencera le choix du dépôt ou de l'archive de données.

Si vous avez besoin d'un archivage à long terme pour votre jeu de données, choisissez un dépôt capable d'assurer la préservation. La préservation numérique peut être coûteuse et longue; certaines données ne peuvent ou ne doivent pas être préservées.

Quelles données allez-vous partager publiquement et sous quelle forme les partagerez-vous (par exemple, brutes, traitées, analysées, finales)?

Considérez quels types de données doivent être partagés pour répondre aux exigences de l'établissement ou de financement, et quelles données peuvent être restreintes pour des raisons de confidentialité ou de propriété intellectuelle.

Avez-vous envisagé le type de licence d'utilisateur final à inclure dans vos données?

Utilisez des licences ouvertes pour promouvoir le partage et la réutilisation des données : Les licences déterminent les utilisations qui peuvent être faites de vos données. Ajoutez une copie de votre licence d'utilisateur final avec votre PGD.

En tant que créateur d'un jeu de données (ou de toute autre œuvre universitaire ou créative), vous en détenez généralement les droits d'auteur par défaut. Cependant, le droit d'auteur empêche d'autres chercheurs de réutiliser et de développer votre travail. Les licences Open Data Commons et les licences Creative Commons (*liens en anglais*) sont un moyen gratuit, simple et standardisé d'accorder des autorisations de droits d'auteur et de garantir une attribution correcte (c'est-à-dire une citation). CC-BY est la licence la plus ouverte et permet à d'autres personnes de copier, distribuer, réarranger et développer votre travail — à condition qu'elles vous citent.

Même si vous choisissez de faire entrer vos données dans le domaine public (sans restriction de réutilisation), il est préférable de le rendre explicite en utilisant une licence telle que la CC0 Creative Commons. Il est fortement recommandé de partager vos données ouvertement en utilisant une licence Open Data ou Creative Commons.

Pour en savoir plus sur l'octroi de licences de données, consultez le site du Digital Curation Centre (*lien en anglais*).

Quels outils et stratégies allez-vous utiliser pour promouvoir votre recherche? Comment ferez-vous savoir à la communauté des chercheurs et au public que vos données existent et sont prêtes à être réutilisées?

Partage de données pour la mobilisation des connaissances : L'utilisation des médias sociaux, des bulletins d'information électroniques, des tableaux d'affichage, des affiches, des conférences, des webinaires, des forums de discussion ou des forums spécifiques à une discipline sont de bons moyens de gagner en visibilité, de promouvoir la transparence et d'encourager la découverte et la réutilisation des données.

L'un des meilleurs moyens de diriger d'autres chercheurs vers vos jeux de données déposés est de les citer de la même manière que vous citez d'autres types de publications. Publiez vos données dans un dépôt qui attribuera un identifiant permanent (tel qu'un DOI) à votre jeu de données. Cela garantira un accès stable au jeu de données et le rendra récupérable par divers outils de découverte. Certains dépôts créent également des liens entre les jeux de données et les documents qui leur sont associés, ce qui augmente la visibilité des publications.

Conformité éthique et juridique

Devez-vous vous conformer à des politiques institutionnelles, gouvernementales ou juridiques par rapport aux normes concernant vos données?

Déterminez votre région : Si votre étude s'effectue dans divers établissements ou différentes régions, vous devrez déterminer quelles lois et politiques s'appliqueront à votre recherche.

Le respect des lois en matière de protection de la vie privée et des lois susceptibles d'imposer des restrictions sur le contenu des données doit être discuté avec le responsable de la protection de la vie privée ou le bureau des services de recherche de votre établissement.

Si vous collaborez avec un partenaire de l'Union européenne, par exemple, vous devrez peut-être suivre le Règlement général sur la protection des données (RGPD).

Si vous travaillez avec des données qui sont la propriété des Premières Nations, des Métis ou des Inuits, par exemple, vous devrez travailler avec des protocoles qui garantissent le respect de la vie privée de la communauté et la réduction des dommages causés à celle-ci. Pour plus d'informations sur la souveraineté des données autochtones, consultez les principes PCAP, et dans un contexte mondial, les principes CARE sur la gouvernance autochtone des données (*lien en anglais*).

Dans le cadre de vos recherches, manipulez-vous de l'information sensible ou nominative? Le cas échéant, comment vous assurerez-vous qu'elles restent sécurisées et que leur accès soit réservé aux membres de l'équipe approuvés uniquement?

Obtenez un consentement éclairé avant la collecte de données : L'obtention du consentement approprié des participants à la recherche est une étape importante pour rassurer les comités d'éthique de la recherche que les données puissent être partagées avec des chercheurs en dehors de votre projet. Votre déclaration de consentement éclairé peut préciser certaines conditions clarifiant les utilisations des données par d'autres chercheurs. Par exemple, votre déclaration peut stipuler que les données ne seront partagées qu'à des fins de recherche non lucratives (vous pouvez utiliser CC-BY-NC — la licence Creative Commons non commerciale avec attribution) ou que les données ne seront pas liées à d'autres données permettant d'identifier des personnes. Notez que cet aspect n'est pas couvert par une licence ouverte. Vous pouvez en savoir plus sur la sécurité des données en consultant le site du [UK Data Service](#) (*lien en anglais*).

Informez les participants à votre étude si vous avez l'intention de publier une version anonymisée et dépersonnalisée des données collectées, et qu'en participant, ils acceptent les présentes conditions. Pour voir des exemples de libellé pour le consentement éclairé, consultez la ressource suivante : [Formulation recommandée de l'ICPSR pour le consentement éclairé au partage des données](#) (*lien en anglais*).

Vous devrez peut-être envisager des stratégies pour garantir la réutilisation éthique de votre jeu de données publiées par d'autres chercheurs. Ces stratégies peuvent influencer votre choix d'une licence appropriée; dans certains cas, vous ne pourrez pas utiliser une licence ouverte.

Avant de publier ou de partager un jeu de données, êtes-vous tenu de retirer les données identificatoires (nom, sexe, date de naissance, etc.) conformément aux lois de votre région ou à votre protocole d'éthique? Existe-t-il des restrictions temporelles concernant le moment où les données peuvent être rendues publiques?

Protection de la confidentialité : Les flux de travail scientifiques ouverts préconisent d'avoir des données « aussi ouvertes que possible et aussi restreintes que nécessaire ». Songez à vos réserves en matière de confidentialité de vos données ou à d'autres restrictions d'accès décrites dans votre protocole d'éthique. Si votre établissement ou votre bailleur de fonds régit des directives juridiques ou éthiques sur les informations qui doivent être protégées, prenez un moment pour vérifier si vous avez respecté les conditions de consentement au partage des données. En l'absence de soutien local pour l'anonymisation ou la dépersonnalisation des données, vous pouvez communiquer avec la personne responsable de la gestion de la planification des données de Portage à l'adresse suivante : support@portagenetwork.ca.