Institut Pasteur: Institut Pasteur - Trame de DMP (FR) - Informations générales sur les données

1. Présentation générale des données

Quel est l'objectif de la collecte/génération des données ?

Recommandations:

Expliquez le lien entre les données générées/collectées et les objectifs du projet

Exemple de réponse:

Nous allons acquérir des images microscopiques des infections par le virus de la fièvre jaune pour effectuer une analyse en super résolution de l'infection cellulaire. Des images microscopiques, des vidéos de l'infection en direct ainsi que des images de Western-Blot sont nécessaires pour une compréhension totale et la réalisation du projet.

Combien de jeu(x) de données allez-vous générer durant ce projet ?

Exemple de réponse:

3 jeux de données

Recommandations:

Un jeu de données est un ensemble de données, brutes ou dérivées, rassemblées pour former un ensemble cohérent. On peut considérer que des données qui sont gérées de la même façon (même méthode de traitement, de stockage, de partage...) forment à un seul jeu de données.

Le nombre de jeux de données peut évoluer au cours du projet.

Dupliquez autant de fois que nécessaire la partie "Jeu de données" : si vous avez 3 jeux de données, vous devez compléter 3 fois la partie "Jeu de données". Pour une aide sur la façon de faire dans DMP OPIDoR, consultez notre guide (partie 4.b. en particulier)

Quelle est la nature et le format des données générées/collectées ?

Exemple de réponse:

- Images de microscopie confocale aux formats .tiff et .PNG
- Feuilles de calcul au format Excel et documents texte en .docx.
- Données de cytométrie en flux en format .FCS.
- Profils de chromatogrammes HPLC de purification Akta au format .csv.
- Images de Western-Blot au format .jpeg et .PNG
- Vidéos de microscopie en format .AVI
- Données épidémiologiques stockées sous forme de base de donnée REDCap

Recommandations:

Précisez la nature des données, qu'elles soient générées ou réutilisées : données de génomique, protéomique, glycomiques, lipidomiques, transcriptomiques, données de structures moléculaires, données d'interactions moléculaires, données cellulaires, physiologiques, phénotypiques, cliniques, épidémiologiques...

Pour vous aider, un document présentant des exemples de types de données et de formats dans le domaine biomédical est accessible ici.

Indiquez le volume prévisionnel des données générées pour ce projet

Recommandations:

Au début du projet, indiquez le volume prévisionnel. S'il vous est inconnu, vous pourrez répondre à cette question en fin de projet.

Exemple de réponse:

3 To

Des données existantes seront-elles réutilisées ? Si oui, précisez leur origine.

- Non applicable
- Non
- Oui

Exemple de réponse:

Ex 1 : Aucune donnée ancienne ne sera utilisée dans ce projet. Tous les livrables du projet proviendront des données originales acquises au cours de cette période.

Ex 2 : des données épidémiologiques issues de 2 cohortes au Cambodge et en Papouasie-Nouvelle-Guinée seront réutilisées. Ces données ont été générées dans le cadre du projet X financé par le NIH.

Recommandations:

Avant de générer des données, il est recommandé de vérifier s'il n'est pas possible de réutiliser des jeux de données produits par d'autres scientifiques. Pour une aide sur les différentes sources à consulter pour trouver des données, consultez notre fiche.

Avant de réutiliser des données, vous devez vous assurer que vous avez le droit de les réutiliser :
- En vérifiant qu'elles ne sont pas protégées par une réglementation nationale ou internationale, un copyright ou un droit de propriété intellectuelle
- En vérifiant que les personnes sont informées de la réutilisation de leurs données personnelles.

Pour vérifier que la réutilisation se fait bien dans le respect du cadre juridique, consultez ce logigramme

En cas de doute, vous devez vous rapprocher du centre de recherche translationnelle (CRT) pour les données concernant la personne humaine (crt-guichetunique@pasteur.fr), ou de la Direction Juridique pour les autres types de données.

2. Ressources nécessaires à gestion des données

De quelles ressources matérielles avez-vous besoin pour la gestion de vos données ?

Exemple de réponse:

Un espace de stockage supplémentaire sera nécessaire. De plus, une base de données devra être mise en place pour gérer les données de spectrométrie au cours du projet.

Recommandations:

Des ressources matérielles peuvent être nécessaires pour la collecte, le stockage, l'analyse et le transfert de vos données. Par exemple, des serveurs de stockage, des ordinateurs, tablettes, téléphones, écrans de sécurité…

Consultez notre fiche pratique pour un résumé des différents outils et infrastructures disponibles à l'Institut Pasteur pour la gestion, le stockage et le partage des données.

Qui assure la gestion opérationnelle des données au cours du projet de recherche ?

Recommandations:

Selon la complexité de votre projet, vous pouvez :
- soit indiquer le rôle de chaque personne individuellement (comme dans l'exemple) ;
- soit indiquer plus généralement quelle équipe de recherche est responsable de la gestion de quel type de données.

Exemple de réponse:

- A est responsable de la collecte, traitement et analyse des données
- B est responsable de la génération des métadonnées et de la documentation accompagnant les données
- C est responsable du stockage des données
- D est responsable de l'archivage et du partage des données

Quelles formations ou quel accompagnement vous semblent nécessaires pour vous aider à la gestion de vos données ?

Recommandations:

Indiquez si une formation est nécessaire (rédaction du PGD, génération des métadonnées, génération des données…). Indiquez également tous les documents ou supports qui vous seraient utiles pour la gestion de vos données...

Exemple de réponse:

Ex 1 : Aucune formation particulière ne sera nécessaire pour conserver et gérer les données générées dans le cadre de ce projet.

Ex 2 : Le responsable de projet aimerait des conseils juridiques et organisationnels sur les sujets suivants : les données à caractère personnel et les licences de réutilisation des données. L'équipe (5 personnes) aura également besoin d'une formation sur des questions techniques : métadonnées, standards de métadonnées et archivage.

Quel budget avez-vous prévu pour la gestion de vos données ? Comment comptez-vous financer ces coûts ?

Recommandations:

Les coûts liés à la gestion des données concernent principalement les ressources matérielles (serveur de stockage, logiciel d'analyse, stockage dans un entrepôt...) et les ressources humaines (embauche d'un data manager par exemple). Au début du projet, indiquez le budget prévisionnel. S'il vous est inconnu, vous pourrez répondre à cette question en fin de projet.

Exemple de réponse:

Ex 1 : Un budget de XXX euros a été prévu pour le recrutement d'un data manager. Les coûts seront financés par la Commission Européenne.
Ex 2 : L'institut et le laboratoire disposent d'un budget dédié pour assurer le stockage et la maintenance des données en interne.

3. Aspects juridiques, éthiques et sécurité

Votre projet comporte-t-il des données à caractère personnel ?

- Non applicable
- Non
- Oui

Exemple de réponse:

Oui, le projet comporte des données à caractère personnel. Les démarches légales pour pouvoir gérer ce type de données ont été effectuées.

Recommandations:

Si vous répondez "oui" à l'une des 2 questions ci-dessous, alors votre projet comporte des données à caractère personnel :
- Est-ce que tout ou partie des données permettent d'identifier une personne humaine ?
(Ex : nom, photo, adresse…)
- Est-ce que tout ou partie des données permettraient d'identifier une personne humaine si on les associait avec d'autres informations détenues par vous ou un tiers ?
(Ex : un numéro d'identifiant, un code avec une table de correspondance détenue par un tiers, des données de localisation, une caractéristique physique, physiologique ou génétique particulière, des éléments propres à une situation psychique, économique,

culturelle ou sociale….)

Si votre projet comporte des données à caractère personnel, vous devez contacter le Délégué à la Protection des Données (dpo@pasteur.fr) ou le pôle de coordination de la recherche clinique (crt-guichetunique@pasteur.fr)

Vous n'avez pas besoin de détailler les démarches entreprises dans la réponse à cette question. Indiquez simplement que les mesures pour gérer ce type de données ont été effectuées. Si cela est pertinent, incluez les références aux livrables éthiques et au chapitre d'éthique de votre projet de recherche.

Votre projet comporte-t-il d'autres données soumises à une obligation contractuelle, réglementaire ou légale ? Si oui, de quel type ?

- Non
- Non applicable
- Oui

Exemple de réponse:

Ex 1 : Le projet comporte des données liées à un contrat avec un industriel qui ne pourront pas être rendues librement accessibles.

Ex 2 : Une demande de dépôt de brevet sera envisagée au cours du projet. Les données ne pourront pas être diffusées avant le dépôt de la demande de brevet.

Recommandations:

Pour vous aider à déterminer si votre projet comporte des données soumises à une obligation réglementaire, contractuelle ou légale, consultez nos deux logigrammes :

Logigramme - Questions juridiques liées à la réutilisation des données

Logigramme - Questions juridiques liées à la diffusion des données

Certaines données doivent-elles rester confidentielles au cours de votre projet ? Si oui, précisez quels types de données sont concernés et auprès de qui ces données peuvent être diffusées.

Exemple de réponse:

Ex 1 : Les données brutes sont des données à caractère personnel qui doivent rester confidentielles et uniquement accessibles aux chercheurs du projet à l'Institut Pasteur. Les résultats intermédiaires pseudonymisés seront accessibles plus largement aux chercheurs du projet (projet multi-partenaires).

Ex 2 : Les données relatives à XXX feront l'objet d'une demande de brevet. Ces données resteront confidentielles avant le dépôt de la demande de brevet. Elles ne seront accessibles qu'aux membres du projet.

Recommandations:

Préciser notamment si certains types de données doivent rester confidentiels car soumis à une obligation contractuelle, réglementaire ou légale. Indiquer si l'accès à ces données doit être restreint (aux pasteuriens et Orex, aux chercheurs du projet, à certaines personnes du projet...).

En particulier, si vous prévoyez de protéger une invention et de déposer une demande de brevet, les données doivent rester confidentielles avant le dépôt de la demande de brevet. Pour toute information complémentaire, n'hésitez-pas à contacter le Service des Brevets et Inventions : sbi@pasteur.fr

Quelles sont les mesures de sécurité mises en œuvre pour le stockage des données pendant le projet ?

Exemple de réponse:

Des mesures de sécurité adaptées seront mises en place pour les données cliniques : données stockées sur des serveurs internes sécurisés à l'Institut Pasteur pendant le projet, dans un dossier accessible uniquement aux chercheurs du projet.

Recommandations:

Les mesures à mettre en place pour sécuriser les données sont dépendantes du niveau de sensibilité des données. Pour toute aide à ce sujet, n'hésitez pas à contacter rssi@pasteur.fr.

Pour déterminer les mesures de sécurité à mettre en place, se référer à la directive de classification des données.

Quelles sont les mesures de sécurité mises en œuvre pour la collecte et l'échange de données ?

Exemple de réponse:

Ex 1 : L'échange de données se fait via Drive Pasteur, espace sécurisé et encrypté de stockage et de partage de fichiers.

Ex 2 : Notre projet ne comporte pas de collecte et d'échange de données. Aucune mesure de sécurité n'est nécessaire.

Recommandations:

Consultez ce document pour une synthèse des moyens de transfert de données depuis et à destination du système d'Information de l'Institut Pasteur

4. Gestion des données au cours du projet

Quel est le lieu de stockage de vos données au cours du projet ?

Exemple de réponse:

Ex 1 : Le projet ne comportant que des partenaires de l'Institut Pasteur Paris, un dossier projet commun a été créé sur les serveurs de stockage de l'Institut Pasteur (espace projet Gaïa).

Ex 2 : Chaque partenaire stocke les données qu'il produit sur ses propres serveurs. Un dossier Drive Pasteur a été créé pour permettre l'échange de données entre partenaires.

Recommandations:

Indiquez si vos données sont stockées :
- Sur votre ordinateur
- Sur un serveur propre à de l'unité
- Sur un espace de stockage partagé fourni par la DSI

Attention, vous ne devez surtout pas stocker vos données sur un espace de stockage sur internet (Dropbox, Google drive, one drive, etc.) car ces espaces ne sont pas sécurisés.

Utilisez-vous un plan de classement pour gérer vos fichiers de données ? Indiquez brièvement comment il est organisé.

- Oui
- Non applicable
- Non

Exemple de réponse:

Oui, un plan de classement a été créé pour l'espace de stockage commun à tous les partenaires du projet. Ce plan de classement est organisé par méthode de collecte des données (microscopie, phénotypage, séquençage...) puis de façon chronologique. Les données brutes et les données traitées sont stockées dans des dossiers différents.

Ci-dessous un aperçu du plan de classement :
I. Méthode de collecte 1 (ex : microscopie)
I.1 Date de la première collecte (ex : 2021-01-12)
I.1.1. Etape de traitement 1 (ex : quality check)
I.1.2. Etape de traitement 2 (ex : données brutes)
I.1.3. Etape de traitement 3 (ex : données analysées)
I.2. Date de la deuxième collecte (ex : 2021-01-19)
II. Méthode de collecte 2 (ex : séquençage)

Recommandations:

Le plan de classement correspond à l'arborescence de dossiers mise en place pour classer les fichiers de données. Cette organisation a pour objectif de permettre à n'importe quel collaborateur de repérer et retrouver facilement et rapidement les données dont il a besoin.

Consultez notre fiche pratique pour vous aider à mettre en place un plan de classement.

Pour les projets de recherche sur la personne humaine, nous vous conseillons de suivre le plan de classement proposé par le pôle de coordination de la recherche clinique. Pour utiliser ce plan de classement, contactez crt-guichetunique@pasteur.fr

Quelles règles de nommage utilisez-vous pour vos données ? Quelles règles utilisez-vous pour un versionnage explicite ?

Exemple de réponse:

Chaque fichier est nommé avec : date et nom du gène/protéine. Toutes les informations sont séparées par un underscore. Pour les fichiers traités et analysés, le nom est suivi de la date d'analyse et de la version du document. Les noms des documents textuels sont suivis de la date et la version ainsi que les initiales des auteurs tout au long de la révision.

Recommandations:

Voir la page rédigée par le service des archives

Les règles pour un bon nommage sont :
- un nom succinct (maximum 30/40 caractères)
- un nom informatif : sujet_TypeDoc_date_version
- un nom interopérable : pas d'espace (underscore uniquement), pas de ponctuation, pas de caractères spéciaux, date écrites de la façon suivante : AAAAMMJJ

Quelles sont les mesures mises en place pour garantir la qualité des données ?

Exemple de réponse:

Pour garantir la qualité des données, différentes mesures ont été mises en place :
- Répétition des expériences de manière indépendante (minimum de trois répétitions sur trois jours différents)
- Standardisation de la collecte des données (tous les animaux élevés dans les mêmes conditions, contrôle de la température, mêmes conditions de stimulation)
- Revue régulière des données avec le PI

Recommandations:

Expliquer comment la qualité de la collecte des données sera contrôlée et documentée. 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 ou la revue par les pairs.

5. Sélection des données pour l'archivage à long terme

Existe-t-il des contraintes réglementaires de conservation de vos données ? Si oui, lesquelles ?

- Non
- Oui
- Non applicable

Exemple de réponse:

Ex 1 : Pas de contrainte réglementaire particulière
Ex 2 : Les contraintes de conservation ont été définies au moment de la définition du protocole. Les données incluant des données à caractère personnel seront supprimées après la publication du dernier article lié à ce projet.

Recommandations:

Des contraintes réglementaires existent essentiellement dans le cas de recherche sur la personne humaine ou utilisant des données de santé. Dans ce cas, les durées de conservation des données doivent être définies dès la définition du protocole.

Pour connaître les contraintes de conservation des données issues de recherche sur la personne humaine, consultez notre fiche pratique.

Quels sont les jeux de données ayant une valeur à long-terme et qui devraient être conservés ? Quels sont les jeux de données à détruire ?

Exemple de réponse:

Ex 1 : Les jeux de données 1 et 2 doivent être conservées en raison des difficultés de reproductibilité et du temps nécessaire à leur régénération. Leur conservation est essentielle pour assurer la reproductibilité des résultats présentés dans les publications et pour pouvoir les comparer aux données qui seront générées ultérieurement.
Ex 2 : les données brutes de séquençage (jeu de données 3) seront supprimées après leur dépôt sur GenBank, afin de gagner de l'espace de stockage.
Ex 3 : Le jeu de données 4 inclut des données à caractère personnel. Il sera supprimé après la publication du dernier article lié à ce projet.

Sur quelle(s) plateforme(s) ou dans quel(s) entrepôt(s) les jeux de données à conserver seront-ils archivés sur le long terme (après la fin du projet) ?

Exemple de réponse:

Ex 1 : A la fin du projet, le jeu de données 1 sera transféré sur l'entrepôt ZENODO qui assure un archivage durable des données de recherche finales.

Ex 2 : Le jeu de données 2 comporte des données sensibles, il ne peut pas être mis à disposition sur un entrepôt externe à l'Institut Pasteur. Il sera donc conservé sur les serveurs de l'Institut Pasteur.

Recommandations:

Les données non sensibles peuvent être conservées sur le long terme sur un entrepôt de données et/ou sur les serveurs de l'Institut Pasteur.
Les données sensibles doivent être conservées sur les serveurs sécurisés de l'Institut Pasteur. Consultez notre fiche pratique pour des conseils d'organisation des espaces sur les serveurs de l'Institut Pasteur

Précisez les formats choisis pour l'archivage.

Exemple de réponse:

Les formats PDF, TIFF, PNG et CSV seront utilisés.

Recommandations:

Choisissez un format ouvert et stable dans le temps si possible. Evitez les formats propriétaires ou les formats qui dépendent de l'environnement technologique.
Voir la fiche pratique précisant la différence entre format ouvert et format fermé.

Pendant combien de temps les données seront-elles conservées ?

Exemple de réponse:

Ex 1 : Les jeux de donnés seront conservés pendant la durée maximale permise par l'entrepôt. Pour Zenodo, cela correspond à la durée de vie du laboratoire hôte, le CERN, qui dispose actuellement d'un programme expérimental défini pour les 20 prochaines années au moins.
Ex 2 : Les données seront conservées pour une durée illimitée tant que l'espace alloué au sein de l'Institut Pasteur est disponible.

Recommandations:

S'il existe une durée légale de conservation des données, citez la réglementation applicable. Si vous considérez que les données doivent être conservées pendant une durée plus longue que la durée légale, justifiez-le.
S'il n'y a pas de réglementation mais que vous pensez que vos données ont une valeur à long-terme, indiquez pendant combien de temps les données seront conservées.

Quel est le volume final de données archivées ?

Exemple de réponse:

2 To

Recommandations:

Au début du projet, indiquez le volume prévisionnel. S'il vous est inconnu, vous pourrez répondre à cette question en fin de projet.

Si une conservation à long terme est requise, comment comptez-vous couvrir ces coûts ?

Exemple de réponse:

Les coûts associés à la conservation à long terme seront pris en charge par l'Institut Pasteur.

Recommandations:

La conservation à long terme des données après la fin du projet peut engendrer des coûts (espaces de stockage principalement)

qui doivent être anticipés pendant le projet.

Institut Pasteur: Institut Pasteur - Trame de DMP (FR) - Jeux de données

1. Description des données

Titre du jeu de données

Qui est le fournisseur ou le producteur des données ?

Exemple de réponse:

Ce jeu de données est généré par le partenaire 1 du projet

Recommandations:

Cette question est importante dans le cas de projets collaboratifs. Vous devez indiquer ici lequel des partenaires va produire ou fournir le jeu de données.

Quelle est la nature et le format des données de ce jeu de données ?

Exemple de réponse:

Ce jeu de données contient des données épidémiologiques sous forme de base de données REDCap et de fichiers Excel et CSV.

Recommandations:

Pour vous aider, un document présentant des exemples de types de données et de formats dans le domaine biomédical est accessible ici.

Décrivez de façon plus détaillée les données de ce jeu de données

Exemple de réponse:

Ce jeu de données comprend des images de cerveau (immunofluorescence et RNAscope) de modèles animaux de souris pour différents gènes de surdité.

Décrivez la méthode de collecte et/ou de génération de ces données

Exemple de réponse:

Les données sont générées par microscopie confocale et analysées par les logiciel ImageJ et NeuroInfo. GraphPad sera utilisé pour l'analyse statistique. Excel sera utilisé pour la gestion des données, les représentations graphiques et l'analyse des données.

Recommandations:

Indiquez comment les données sont générées ou collectées : données générées par une machine, enquête, observation, simulation, analyse... S'il s'agit de données réutilisées, non générées ou collectées pendant le projet, indiquez leur source (autre laboratoire, base de données en ligne...)

Décrivez votre jeu de données avec des mots-clés

Exemple de réponse:

Larve de drosophile, classification comportementale, séquences, interactions compétitives, substrats neuronaux

Recommandations:

Nous vous conseillons de décrire le jeu de données par 3 mots-clés au minimum. Un jeu de données décrit précisément sera plus facilement trouvé et donc réutilisé.

Indiquez l'URL ou l'identifiant pérenne permettant d'accéder à votre jeu de données

Recommandations:

Certains entrepôts de données attribuent des identifiants pérennes aux jeux de données. Si c'est le cas, indiquez l'identifiant pérenne ici. Sinon, indiquez l'URL permettant d'accéder au jeu de données.

Exemples d'identifiants pérennes : Handle, DOI (Digital Object Identifier), Ark...

Exemple de réponse:

Ex 1 : non disponible à ce stade du projet, l'identifiant sera indiqué dès publication du jeu de données.

Ex 2 : https://doi.org/10.17867/10000105

Quel est volume de données dans ce jeu de données ?

Exemple de réponse:

1 To

2. Rendre les données librement accessibles

Ce jeu de données sera-t-il librement disponible ?

- Oui
- Non applicable
- Non

Recommandations:

Indiquez si le jeu de données finalisé sera librement disponible pour la communauté scientifique (en fin de projet, au moment de

la publication de l'article associé, etc...).

Si ce n'est pas le cas (jeu de données accessible mais avec certaines restrictions, jeu de données non accessible…), expliquez pourquoi dans la question suivante.

Exemple de réponse:

Ex 1 : Ce jeu de données sera rendu librement disponible au moment de la prépublication de l'article associé.

Ex 2 : Non, ce jeu de données ne sera pas librement disponible car il comporte des données à caractère personnel.

Si ce jeu de données ne peut pas être librement diffusé, expliquez pourquoi.

Exemple de réponse:

Ex 1 : Ce jeu de données contient des données à caractère personnel non anonymisées et non pseudonymisées. Il ne peut donc pas être rendu public librement.

Ex 2 : Ce jeu de données a été produit en collaboration avec une entreprise privée. Le contrat avec cette entreprise prévoit que les données ne pourront pas être rendues publiques.

Ex 3 : Ce jeu de données fera l'objet d'une demande de brevet. Il ne peut donc pas être librement diffusé.

Recommandations:

Certaines données ne peuvent pas être librement diffusées car il s'agit de données soumises à une obligation réglementaire, contractuelle ou légale.

Consultez le logigramme "Questions juridiques liées à la diffusion des données" pour en savoir plus

Quel entrepôt de données avez-vous choisi pour stocker et rendre accessible ce jeu de données ?

Exemple de réponse:

Ex 1 : Ce jeu de données sera rendu librement disponible via l'entrepôt PRIDE/Zenodo/ENA/...

Ex 2 : Ce jeu de données sera déposé sur l'entrepôt European Genome-phenome Archive (EGA), entrepôt permettant de contrôler l'accès aux données sur la personne humaine.

Ex 3 : En raison de contraintes juridiques, ce jeu de données ne sera pas déposé sur un entrepôt de données, il sera conservé sur les serveurs de l'Institut Pasteur.

Recommandations:

Les entrepôts de données sont la meilleure solution pour partager ses données à un large public. De plus, certains entrepôts permettent de contrôler l'accès aux données : partager ses données via un entrepôt ne signifie pas que les données seront accessibles à tous sans restriction.

Pour vous aider à trouver l'entrepôt adapté à vos besoins, consultez notre fiche pratique.

Précisez comment l'accès à ce jeu de données sera fourni au cas où des restrictions s'appliquent.

Exemple de réponse:

Ex 1 : Toute demande d'accès aux données déposées sur l'entrepôt European Genome-phenome Archive (EGA) sera vérifiée par le Data Access Committee du projet. Pour toute demande d'accès, contacter xxx@pasteur.fr

Ex 2 : Le jeu de données est déposé sur Zenodo mais en accès restreint. Les conditions d'accès sont précisées sur Zenodo : la personne qui souhaite télécharger le jeu de données doit au préalable expliquer comment elle a l'intention de l'utiliser. Sur la base de cette justification, il sera décidé d'accorder ou de refuser l'accès.

Ex 3 : Le jeu de données est conservé sur les serveurs de l'Institut Pasteur et accessible sur demande à xxx@pasteur.fr

Recommandations:

Si le jeu de données est stocké sur un entrepôt mais non accessible librement, indiquez comment il est possible d'accéder au jeu de données : accès sur demande, sous réserve de l'approbation d'un comité scientifique....

Si le jeu de données est conservé sur les serveurs de l'Institut Pasteur, indiquez qui contacter pour demander l'accès aux données, quelles sont les conditions d'accès...

Quels logiciels sont nécessaires pour visualiser ou accéder aux données ? Fournissez-vous une documentation ou le code open source du logiciel ?

Recommandations:

Indiquez quel logiciel vous utilisez pour visualiser, lire ou analyser les données.
Si vous avez développé un logiciel spécifique, indiquez où est déposé le code source ou comment y accéder.

Exemple de réponse:

Ex 1 : L'accès aux données nécessite un logiciel développé par notre unité. Pour rendre nos données accessibles, nous fournissons le code open source de ce logiciel.

Ex 2 : Logiciels utilisés régulièrement : ImageJ, GraphPad

3. Rendre les données trouvables

Ce jeu de données est-il identifié par un identifiant pérenne unique tel que le DOI (Digital Object Identifier) ? Sinon, décrivez comment ce jeu de données et les données sont identifiés.

- Non
- Non applicable
- Oui

Exemple de réponse:

Ex 1 : Oui, ce jeu de données est identifié par un DOI fourni par l'entrepôt Zenodo.

Ex 2 : Ce jeu de données est identifié par un Accession Number (identifiant unique mais non pérenne) fourni par l'entrepôt Sequence Read Archive (SRA).

Ex 3 : Non applicable, ce jeu de donnée n'est pas déposé sur un entrepôt de données, il ne comporte donc pas d'identifiant.

Recommandations:

Le type d'identifiant dépend essentiellement de l'entrepôt choisi pour déposer vos données. La plupart des entrepôts attribuent des identifiants uniques aux jeux de données, mais ils ne sont pas nécessairement pérennes.

Pour vérifier si l'entrepôt choisi attribue des identifiants pérennes, cherchez l'entrepôt dans Re3data et regardez si la 4e icône est bleue. Sous l'onglet 'standards', le type d'identifiant est indiqué en premier.

Quels standards de métadonnées utilisez-vous ? Si vous n'utilisez pas de standard de métadonnées, précisez quel(s) type(s) de métadonnées seront créés et comment.

Exemple de réponse:

Ex 1 : Nous prévoyons de déposer ce jeu de données sur l'entrepôt PRIDE, entrepôt utilisant le standard de métadonnées "Minimum Information About a Proteomics Experiment (MIAPE)". Nous décrirons donc ce jeu de données en suivant ce standard, dès la génération des données.

Ex 2 : Nous n'avons pas encore défini l'entrepôt dans lequel les données seront déposées. Nous suivrons le standard de métadonnées ou les règles de l'entrepôt choisi pour stocker les données.

Ex 3 : Etant donné qu'aucun standard disciplinaire n'existe dans notre domaine et que ce jeu de données ne sera pas déposé dans un entrepôt, nous avons défini nos propres métadonnées qui seront collectées dans un fichier Excel et associées aux données : XXXX

Recommandations:

Indiquez quelles métadonnées sont associées aux données de façon à ce que les données soient décrites précisément et donc réutilisables. Les métadonnées sont des informations structurées qui décrivent les données. Elles peuvent être généralistes (ex: auteur, format,date de création...) ou plus scientifiques (ex: organisme, échantillon, allèle, pathologie...).

Pour vous aider à répondre à cette question, consultez notre document sur les métadonnées et standards de métadonnées.

Ce jeu de données est-il décrit par des mots-clés permettant de le retrouver facilement ?

- Oui
- Non applicable
- Non

Exemple de réponse:

Oui, ce jeu de données est décrit par 3 mots-clés minimum

Fournissez-vous une documentation complémentaire pour décrire plus précisément les données ?

- Non applicable
- Non
- Oui

Recommandations:

La documentation correspond à un texte qui décrit les données, les contextualise et donne toute information nécessaire à leur compréhension. Il peut s'agir d'un fichier README associé à chaque jeu de données, du protocole de recherche ou de tout autre document utile à la compréhension des données.

Exemple de réponse:

Oui, un fichier README est associé au jeu de données pour le contextualiser et récapituler les analyses effectuées. Ce fichier est rédigé par l'expérimentateur au moment de la génération des données et sera partagé en même temps que les fichiers de données.

4. Rendre les données interopérables

Les données de ce jeu de données sont-elles interopérables du point de vue technique ?

- Oui
- Non applicable
- Non

Exemple de réponse:

Oui, les photographies de microscopie sont au format PNG. Les tableaux de résultats accompagnant les photographies sont au format CSV. Les formats PNG et CSV sont des formats ouverts et donc interopérables.

Recommandations:

Des données sont interopérables si elles peuvent être facilement combinées avec d'autres données. Du point de vue technique, cela dépend essentiellement du format dans lequel elles sont enregistrées. Il est recommandé d'utiliser un format ouvert, largement diffusé et que de nombreux logiciels peuvent exploiter.

Consultez notre fiche pratique pour une explication sur la différence entre formats ouverts et fermés.

Sinon, quelles méthodologies appliquerez-vous pour rendre vos données interopérables ?

Recommandations:

Si vos données sont dans un format propriétaire, vous pouvez : si possible les transformer dans un format ouvert et interopérable, sinon indiquer dans les métadonnées associées au jeu de données le nom et la version du logiciel nécessaire pour lire les données.

Si vous avez développé un logiciel pour produire ou analyser les données, il est conseillé de le partager.

Exemple de réponse:

Nos données sont dans un format lisible uniquement avec un logiciel développé par notre service. Toutefois, nous fournissons le code open source du logiciel nécessaire pour accéder aux données (documentation annexe).

Précisez si vous utilisez du vocabulaire standard dans votre jeu de données pour permettre une interopérabilité sémantique interdisciplinaire. Dans le cas contraire, fournirez-vous un alignement avec les ontologies les plus fréquemment utilisées ?

Exemple de réponse:

Ex 1 : Comme notre projet concerne des produits médicaux à usage humain, nous avons utilisé le medDRA (Medical Dictionary for Regulatory Activities) pour décrire nos données.

Ex 2 : Nous n'avons pas utilisé d'ontologie spécifique mais nous utilisons un vocabulaire uniforme dans l'ensemble du jeu de données. Le vocabulaire utilisé est basé sur des termes couramment utilisés dans les sciences du vivant, les termes spécifiques sont précisés dans la documentation associée au jeu de données.

Ex 3 : Le jeu de données sera déposé dans la base de données SRA (Sequence Read Archive) qui permet l'alignement avec plusieurs ontologies : Disease Ontology, Cell Ontology, Uberon, Experimental Factor Ontology, Cellosaurus

Recommandations:

Une ontologie définit un vocabulaire commun pour les chercheurs qui ont besoin de partager l'information dans un domaine. Elle inclut des définitions lisibles en machine des concepts de base de ce domaine et de leurs relations.

Si vous connaissez l'entrepôt de données dans lequel vous allez déposer le jeu de données, vous pouvez chercher quelle ontologie est utilisée pour la description des données et l'indiquer dans votre réponse.

5. Accroître la réutilisation des données

A la fin du projet, ce jeu de données pourra-t-il être réutilisé par des parties tierces ? Si la réutilisation est restreinte, expliquez pourquoi.

Recommandations:

Si un jeu de données qui peut intéresser le public est mis à disposition de tous, n'hésitez pas à contacter la Direction de la Communication, ce qui permettra de valoriser les recherches de l'Institut Pasteur.

La réutilisation des jeux de données ayant fait l'objet d'une demande de brevet est restreinte. En effet, même si les données sont rendues publiques après le dépôt de la demande de brevet, elles ne peuvent pas être réutilisées par des parties tierces sans une licence d'exploitation de la demande de brevet ou du brevet. Pour toute information complémentaire, n'hésitez-pas à contacter le service des Brevets et Inventions : sbi@pasteur.fr

Exemple de réponse:

Ex 1 : Une fois publiées, les données de ce jeu de données pourront être réutilisées par l'ensemble de la communauté scientifique, avec la restriction de ne pas utiliser les données dans un but commercial (une licence CC-BY-NC sera associée au jeu de données)

Ex 2 : Le jeu de données fera l'objet d'une demande de brevet. Les données ne pourront donc pas être réutilisées sans l'accord du titulaire du brevet : toute réutilisation nécessite une licence d'exploitation de la demande de brevet ou du brevet. Par exception, les données pourront uniquement être réutilisées pour vérifier expérimentalement que le brevet fonctionne (exemption de recherche).

Quelle licence sera attribuée à ce jeu de données afin de permettre la réutilisation la plus large possible ?

Recommandations:

Une licence de diffusion est un instrument juridique qui permet au titulaire des droits sur les données d'accorder à l'avance aux utilisateurs certains droits d'utilisation des données. Une licence peut également comporter des restrictions d'usage (pas d'utilisation commerciale par exemple). Pour un usage international, nous vous conseillons les licences Creative Commons.

Consultez cet outil en ligne pour vous aider à choisir votre licence Creative Commons.

Exemple de réponse:

Ex 1 : L'entrepôt Zenodo choisi pour la publication de ce jeu de données permet de choisir la licence à associer aux données. En conséquence, ce jeu de données sera rendu librement disponible sous licence CC-BY-NC (https://creativecommons.org/licenses/by-nc/4.0/).

Ex 2 : Ce jeu de données sera déposé sur l'entrepôt OpenNeuro qui met à disposition les données sous licence CC0 (licence imposée).

Ex 3 : Ce jeu de données ne sera pas rendu librement disponible, aucune licence n'y est donc associée. L'accès au jeu de données sera fourni sur demande et les conditions d'utilisation seront définies au moment du partage.

Quand le jeu de données sera-t-il accessible pour la réutilisation ? Le cas échéant, précisez pour quelle raison et pendant combien de temps un embargo est nécessaire.

Exemple de réponse:

Les données pourront être réutilisées dans un an, après la soumission et la publication de l'article.

Recommandations:

Vous pouvez choisir de ne pas permettre la réutilisation de vos données pendant une certaine durée (embargo). Par exemple, si

vous souhaitez mener d'autres recherches à partir de ces données.

Précisez la durée pendant laquelle le jeu de données sera réutilisable

Exemple de réponse:

Les données seront conservées pendant la durée maximale permise par la capacité de stockage de l'Institut Pasteur.