Gestion des données : DMP OPIDoR et le cycle de vie des données

Paulette Lieby 0000-0002-9289-9652

DMP OPIDoR

Attention : la version présentée ici est celle du 12/11/21 : V2.3.4

Très bientôt (mi-novembre 2021) la version maDMP va être mise en production et sera donc celle avec laquelle vous travaillerez.
Cette version sera présentée dans un des modules de cette formation.

Gestion des données :
DMP OPIDoR, un outil d'aide au PGD

Complément aux modules 2 & 3

DMP OPIDoR https://dmp.opidor.fr/ description des données

DMP OPIDoR https://dmp.opidor.fr/ description des données

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

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

DMP OPIDoR https://dmp.opidor.fr/ description des données -- examples

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

Platform : Illumina

Instrument : Illumina Hiseq 2000

Réutilisation de données déjà produites.

Mise en place d'un workflow d'analyse automatisé et reproductible

ANR

- Expliquer quelles méthodologies ou quels logiciels seront utilisés si de nouvelles données sont recueillies ou produites.
- Enoncer les éventuelles restrictions à la réutilisation des données préexistantes.
- Expliquer comment la provenance des données sera documentée.
- Indiquer brièvement le cas échéant, les raisons pour lesquelles l'utilisation de sources de données existantes a été envisagée mais écartée.

DMP OPIDoR https://dmp.opidor.fr/ description des données -- examples

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

Origine : échantillons humains de différents types (infant oral cavity, blood, feces, cantal, vaginal, skin, milk,...)

Type : lectures (données brutes de séquençage) au format texte

Plateforme : Illumina, Instrument : HiSeq2000

Format : fastq (format standart)

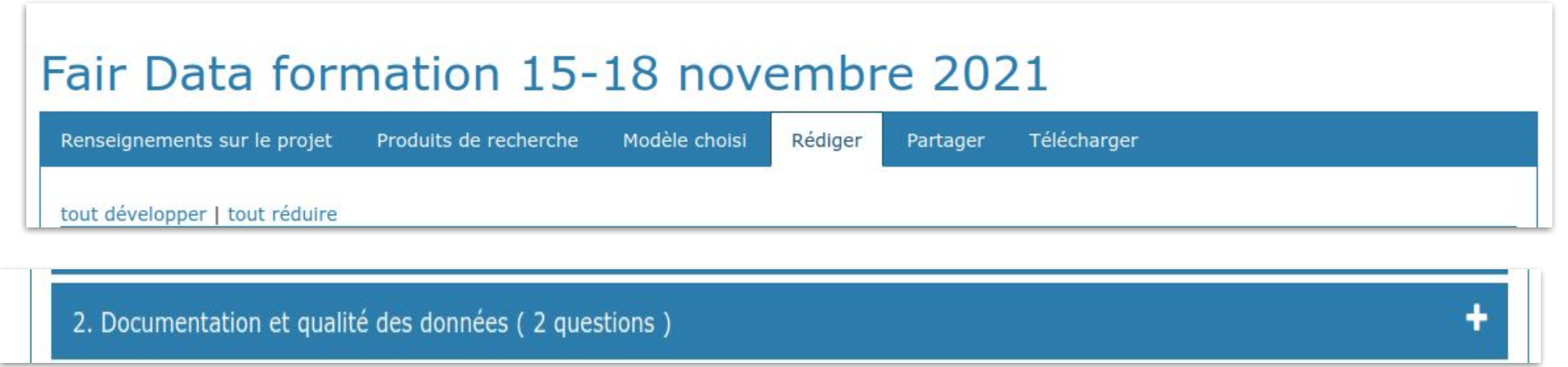
Tableau des échantillons

tax_id	scientific_name	Strain	Origin	Isolation Source	Sa... ye...
1304	Streptococcus salivarius	B35	lot_labo	Human blood	201
1304	Streptococcus salivarius	B50	lot_labo	Human blood	200
1304	Streptococcus salivarius	B57	lot_labo	Human blood	200

ANR

- Donner des détails sur le type de données : par exemple numérique (bases de données, tableurs), textuel (documents), image, audio, vidéo, et/ou médias composites.
- 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").
- Donner 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).

DMP OPIDoR https://dmp.opidor.fr/ métadonnées et documentation



DMP OPIDoR https://dmp.opidor.fr/ métadonnées et documentation

2a. Quelles métadonnées et quelle documentation (par exemple méthodologie de collecte et mode d'organisation des données) accompagneront les données ?

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

DMP OPIDoR https://dmp.opidor.fr/ métadonnées et documentation -- examples

2a. Quelles métadonnées et quelle documentation (par exemple méthodologie de collecte et mode d'organisation des données) accompagneront les données ?

Source des métadonnées : Checklistes ENA (https://www.ebi.ac.uk/ena/browser/checklists)

Rechercher la plus adaptée à ce PGD

Standard : XML

ANR

- Indiquer quelles métadonnées seront fournies pour aider à la recherche et à l'identification des données.
- Indiquer quels standards de métadonnées seront utilisés (par exemple DDI, TEI, EML, MARC, CMDI).
- Utiliser les standards de métadonnées des communautés scientifiques lorsque ceux-ci existent.
- Indiquer comment les données seront organisées au cours du projet, en mentionnant par exemple les conventions de nommage, le contrôle de version et les structures des dossiers. Des données bien classées et gérées de façon cohérente seront plus faciles à retrouver, à comprendre et à réutiliser.

- Penser 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.
- Tenir 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.

DMP OPIDoR https://dmp.opidor.fr/ métadonnées et documentation -- examples

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

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

Contrôle de la qualité des données avec FastQC

- *Version* : 0.11.9
- *Commande d'installation conda* : conda install -c bioconda fastqc=0.11.9
- *Documentation* : https://anaconda.org/bioconda/fastqc

Génération d'un rapport de qualité avec MultiQC

- *Version* : 1.9
- *Commande d'installation conda* : conda install -c bioconda multiqc=1.9
- *Documentation* : https://anaconda.org/bioconda/multiqc

ANR

- 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.

DMP OPIDoR https://dmp.opidor.fr/ stockage et sauvegarde pendant le projet

DMP OPIDoR https://dmp.opidor.fr/ stockage et sauvegarde pendant le projet

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

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

DMP OPIDoR https://dmp.opidor.fr/ stockage et sauvegarde pendant le projet -- examples

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

Espace projet sur le cluster de l'IFB : /shared/projects/monProjet

Cette espace est sauvegardé fréquemment.

ANR

- 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.

DMP OPIDoR https://dmp.opidor.fr/ stockage et sauvegarde pendant le projet -- examples

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

Accès par un compte sécurisé

1. Création d'un compte personnel avec une adresse institutionnelle
2. Validation de la création par des admins
3. Accès au service par une connexion sécurisée en ssh (mot de passe)

ANR

- Expliquer 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.
- Tenir 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 est mise en œuvre.

Merci !
Questions ?

Remerciements : Hélène Chiapello, Thomas Denecker, Frédéric de Lamotte.