PRODIG - UMR: PRODIG- Modèle de DMP

Identifiants du DMP

Identification du document livré - Modèle du DMP

Exemple de réponse:
ANR - Modèle de DMP (français)

Identification du document livré - Historique des mises à jour du DMP
Préciser les dates de transmission au bailleur du DMP aux différentes étapes (en général au début, au milieu, à la fin)

Exemple de réponse:
- Version 3 : 27/04/2020
- Version 2 : 5/12/2019
- Version 1 : 1/3/2018 - date de dépôt effectif à la structure exigeant le DMP en tant que livrable (en général à remettre dans les 6 mois suivant le début du financement, cf. cas ANR et H2020)

Identification du document livré - Date de dernière modification du DMP

Identification du document livré - Créateur(s) du DMP
Prénom Nom, affiliation principale (Institution, Laboratoire, Ville, Pays OU identifiant type ORCID)

Exemple de réponse:
Jean-Yves Menut, CNRS, Prodig, UMR 8586, Paris, France
OU Jean-Yves Menut, ORCID 0000-0001-9010-500X

Identification du document livré - Droits d'auteur

Exemple de réponse:
Les créateurs de ce plan acceptent que tout ou partie du texte de ce plan soit réutilisé et personnalisé si nécessaire pour constituer un autre plan.

Identification du document livré - Contact pour les Données : *Adresse mail du créateur du DMP*

Identification du projet de recherche concerné - Intitulé du projet de recherche : *Titre et acronyme du projet*

Exemple de réponse:
Caractérisation des RIsques de submersion sur des Sites Sensibles, CRISSIS

Identification du projet de recherche concerné - Financeur(s)

Exemple de réponse:
Agence nationale de la recherche (ANR)

Identification du projet de recherche concerné - Référence ou Numéro de subvention du projet ou de la convention de financement

Exemple de réponse:
ANR-29-CE35-0008, AAP ACC 2017

Identification du projet de recherche concerné - Dates de début et de fin du projet : *Dates de début et fin du financement ou de début et fin des travaux de recherche en l'absence de convention de financement*

Identification du projet de recherche concerné - Chercheur(s) Principal(aux) ou coordinateurs scientifiques : *Prénom Nom, (Institution d'affiliation, Laboratoire, Ville, Pays OU identifiant type ORCID)*

Exemple de réponse:
Jean Menut, CNRS, Prodig, UMR 8586, Paris, Paris, France
OU Jean-Yves Menut, ORCID 0000-0001-9010-500X

Identification du projet de recherche concerné - Structures partenaires du projet : *Institution, Laboratoire, Ville, Pays*

Exemple de réponse:
IRD, GRED, Montpellier, France
Université Gaston Berger, Dakar, Sénégal

Recommandations:
Ce champ comporte la liste des établissements et partenaires du projet, notamment ceux impliqués dans la production et la gestion de données. Dans le cas d'un projet associant des partenaires sur d'autres sites ou dans d'autres institutions, les responsabilités entre les partenaires doivent être précisées

Identification du projet de recherche concerné - Thématiques du projet

Exemple de réponse:

Changement climatique / Biotechnologies et Bioressources / Santé

Recommandations:

En référence par exemple aux Domaines ERC :
https://ec.europa.eu/research/participants/data/ref/h2020/other/guides_for_applicants/h2020-guide20-erc-stg-cog_en.pdf

Identification du projet de recherche concerné - Mots-clés du projet : *à compléter d'après les mots-clés utilisés lors du dépôt du projet*

Identification du projet de recherche concerné - Résumé du projet de recherche *(d'après le résumé écrit lors du dépôt du projet +/- actualisation en cours de projet)*

Identification du projet de recherche concerné - Objectifs et attendus du projet *(d'après le résumé écrit lors du dépôt du projet +/- actualisation en cours de projet)*

1. Données collectées et/ou réutilisation de données existantes

1.1. Cas des données recueillies ou produites

Méthodes et outils utilisés pour acquérir et traiter les données. Préciser entre autres le périmètre, l'échelle, la couverture temporelle. Expliquer quelles méthodologies ou quels logiciels sont utilisés pour produire, collecter, pré-traiter les données. Préciser si besoin la fréquence de collecte des données au cours du projet.

Exemple de réponse:

- Jeux de données collectées :

(1) Dispositif statistique expérimental

(2) Calcul d'indice de végétation NDVI / HiSeq 2000 Sequencing System - Illumina de la plateforme de génomique GenoToul (GeT)

(3) Alignements nucléiques sur génome de référence du sorgho

(4) interviews effectuées à Dakar selon une grille d'entretien fixée puis retranscrits

- Passation de questionnaires chaque deux mois, avec suivi longitudinal d'une cohorte de 150 individus ; réponses enregistrées sous forme de base de données anonymisées Epidata, version 3.0.

- Données recueillies sur le dispositif STRADIV (60*9*15m) à Ivory (Madagascar) + 17 parcelles paysannes : ltk, matériel végétal, plan de l'essai, intrants, biomasse (riz, stylo), adventices, faune du sol (monolithes, pit fall trap), vers blancs, rendement estimé (riz, arachide, sorgho), composante du rendement en riz.

- Expérimentation, observation, captation d'images satellitaires…

Recommandations:

Préciser, si besoin, la part des données brutes et dérivées.

Renvoyer éventuellement au protocole de recherche accessible en ligne.

1.2. Cas de l'utilisation de données existantes

Si le projet mobilise des données déjà existantes, préciser lesquelles, leur(s) origine(s) et les modalités de leur intégration aux jeux de données mobilisés par le projet. Si du matériel protégé par des droits spécifiques est utilisé au cours du projet, préciser les éventuelles contraintes d'accès et/ou réutilisation fixées par les licences qui les protègent (type Licences Creative Commons version 4.0 : https://creativecommons.org/licenses/by/4.0/)

Exemple de réponse:

- Déclaration annuelle des données sociales (DADS) : postes et salariés au 1/25, 1996, INSEE (producteur), ADISP (diffuseur)

- Achat de données auprès de l'IGN

- Enquêtes qualitatives réalisées en 2000-2003 dans le cadre du projet ANR « Villes du Sud » (Références, institutions) dans la ville de Mexico, quartier de Cuajmalpa : données qui ont été numérisées et mises à disposition de la communauté scientifique dans l'entrepôt BeQuali. (Références du dépôt) et accessibles sur demande motivée et contrôlée (Références du contrat de réutilisation garantissant notamment l'obligation de confidentialité)

- Données produites lors de 4 enquêtes cas-témoins réalisées par une équipe partenaire du projet depuis 1998 dans le cadre du programme VIH et migrations 1 et 2.

Recommandations:

Préciser si besoin les modalités de réutilisation et les moyens utilisés pour se mettre en conformité avec ces modalités. Vérifier notamment en cas d'agrégation de nouvelles données avec d'anciens jeux de données que la licence de réutilisation permet ce cas de figure.

1.3.1. Types de supports et formats des données collectées ou utilisées

Préciser les formats de production, de traitement ou de conversion des données utilisés au cours du projet ainsi que les outils et logiciels de lecture associés.

Types de supports:

- Image : JPEG, JPG-2000, PNG, TIFF
- Texte : texte brut (TXT), HTML, XML, PDF/A
- Audio : AIFF, WAVE
- Logiciels de compression de fichiers : TAR, GZIP, ZIP
- SIG : Esri Shapefiles
- Données géographiques : KML (.kml), SHP (.shp)
- Bases de données : préférez XML ou CSV aux formats binaires natifs

Exemple de réponse:

- Quantitative survey data files generated will be processed and submitted to the [repository] as SPSS system files with DDI XML documentation. The data will be distributed in several widely used formats, including ASCII, tab-delimited (for use with Excel), SAS, SPSS, and Stata. Documentation will be provided as PDF. Data will be stored as ASCII along with setup files for the statistical software packages. Documentation will be preserved using XML and PDF/A

- Les enregistrements audio et les prises de notes effectués durant les entretiens feront l'objet d'une retranscription (fichier texte avec mise en forme) en français (avec traduction le cas échéant pour les entretiens menés dans une autre langue). Lors de la retranscription, les entretiens seront codés (changement des noms) afin de les pseudonymiser.

Recommandations:

C'est dès cette étape que doit être envisagée la préservation future des données :

Privilégier des formats ouverts, non-propriétaires et formalisés et/ou d'un usage très répandu au sein de votre communauté de recherche pour leur préservation et accessibilité ultérieure. Voir par exemple : https://www.ukdataservice.ac.uk/manage-data/format/recommended-formats

Penser à prévenir l'obsolescence des fichiers et à garantir la meilleure durabilité possible des données. Voir par exemple : https://dmptool.org/general_guidance#file-formats; https://sciencespo.libguides.com/donnees-de-la-recherche/gerer

1.3.2. Volumétrie prévisionnelle des données collectées ou utilisées

Estimation de la taille du/des jeu(x) de données (pouvant être modifiée au cours et en fin de projet, soit au format : [X] Mégaoctets, [n] Téraoctets, soit en utilisant d'autres unités de grandeur (nombre de fichiers, durée des enregistrements, etc)

Exemple de réponse:

Moins de 50 Mo, entre 50 et 100 Mo, 1 G, 80 heures d'entretiens.

1.4. Potentiel(s) de réutilisation des données collectées ou produites

Indiquer quelles pourraient être les réutilisations futures du/des jeu(x) de données utilisées dans le projet, et/ou leurs perspectives d'application ou de développement. S'il existe déjà des données dans ce champ, préciser l'apport spécifique du/des jeu(x) de données produites.
Cette entrée ne préjuge pas de l'ouverture effective au partage des données, abordée infra 5.

Exemple de réponse:

- Comparaison multi-sites,
- Classification et méthodes de mise en relation de données multi-sources,
- Méta-analyses,
- Extraction de connaissances par des méthodes de fouille de textes.

2. Documentation et qualité des données

2.1. Métadonnées et/ou documentation accompagnant les données

Expliquer l'ensemble structuré d'informations décrivant les données pour les rendre consultables, lisibles et interprétables. A minima, un fichier de type « Lisez-moi » doit être prévu pour lister les informations de base de description des données.
Selon les disciplines, les thématiques, les institutions ou selon l'entrepôt où les données vont être stockées, préciser les formats ou standards de description des données utilisés. Si aucun standard (par exemple disciplinaire) n'existe, expliquer et documenter la solution choisie.

Exemple de réponse:

- Les données issues de l'imagerie médicales seront décrites selon la norme et le format d'interopérabilité DICOM.

- Il n'existe pas à notre connaissance de standards de métadonnées pour décrire les données produites mais un modèle sera mis au point et documenté dans le cadre du projet. Les outils et logiciels de gestion développés dans ce contexte seront également documentés et autant que possibles rendus accessible.

- Chaque retranscription d'entretien sera accompagnée d'un rapport d'entretien incluant : le contexte de passation de l'entretien, les caractéristiques sociodémographiques de la ou des personnes interrogées, un résumé et des mots-clés.

Recommandations:

Standards de métadonnées :

- **Dublin Core** : les quinze éléments de base qui décrivent le contenu, la propriété intellectuelle ainsi que les instances particulières (Date, Type, Format, Identifiant) auxquels s'ajoutent une quarantaine d'éléments plus spécifiques qui viennent préciser les éléments simples (date de création ou mise à disposition, format ou taille de la ressource, référence à une ressource apparentée). https://www.enssib.fr/bibliotheque-numerique/documents/1236-presentation-des-standards-le-dublin-core-dc.pdf

- **DDI (Data Documentation Initiative)** pour les Enquêtes qualitatives en sciences humaines et sociales. C'est un standard libre pour documenter et gérer différentes étapes du cycle de vie des données de recherche, comme la conceptualisation, la collecte, le traitement, la distribution, la découverte et l'archivage. https://www.ddialliance.org/

- **DCMS (DataCite Metadata Schema)** : liste des propriétés de métadonnées de base choisies pour une identification précise et cohérente d'une ressource à des fins de citation et de récupération (jeux de données, logiciels...). Éléments obligatoires : DOI, Title, Creator (Producteurs de données), Publisher (Editeur), Publication Year, Resource Type. https://schema.datacite.org/

- **MIAPPE** (Normes de données pour le phénotypage des plantes) https://www.miappe.org/

- **Directive INSPIRE** : elle vise à établir une infrastructure d'information géographique dans la Communauté européenne pour favoriser la protection de l'environnement. Elle comprend un ensemble de métadonnées pour les séries de données géographiques et pour les ensembles de données géographiques. Il y a obligation d'attribuer un mot-clé issu du thésaurus INSPIRE GEMET Spatial Themes. http://www.geoinformations.developpement-durable.gouv.fr/guide-de-saisie-des-elements-de-metadonnees-a2452.html

- **Norme ISO 19115-1:2014** pour l'information géographique : elle définit le schéma requis pour décrire l'information géographique et les services au moyen de métadonnées. Il fournit des informations sur l'identification, l'étendue, la qualité, les aspects spatiaux et temporels, le contenu, la référence spatiale, la représentation, la distribution et les autres propriétés des données et services géographiques numériques. https://www.iso.org/standard/53798.html

Modèles de vocabulaires contrôlés (mots-clés, vocabulaires et ontologies pour la description des contenus) :

- **Noms des pays** : Geonames http://www.geonames.org/
- **Noms géographiques** : Getty Thesaurus of Geographic Names® Online https://www.getty.edu/research/tools/vocabularies/tgn/
- **Agronomie, alimentation, plantes et biodiversité** : http://agroportal.lirmm.fr/
- **Alimentation et agriculture** : Thésaurus Agrovoc (à compléter si nécessaire par des mots-clés spécifiques) http://aims.fao.org/fr/agrovoc
- **Agronomie** : Ontology of agronomic practices, agronomic techniques, and agronomic variables used in agronomic experiments https://www.ebi.ac.uk/ols/ontologies/agro
- **Démographie** : Thésaurus DemoVoc (INED) http://thesaurus.web.ined.fr/navigateur/demovoc/fr/
- **Géographie et environnement** : INSPIRE GEMET Spatial Data Themes: C' est un thésaurus multilingue sur les thèmes de l'environnement de la géographie et d'autres disciplines, servant d'outil d'indexation, de recherche et de contrôle pour le Centre thématique européen sur catalogue des sources de données (ETC / CDS) et l'Agence européenne pour l'environnement (AEE). https://www.eionet.europa.eu/gemet/en/themes
- **Santé, domaine du cancer** : NCI Thesaurus OBO Edition https://www.ebi.ac.uk/ols/ontologies/ncit

2.2. Gestion des fichiers et règles de nommage

Indiquer les règles d'organisation et les conventions de nommage des fichiers mises en place dès le début du projet afin de faciliter la recherche puis la mise à disposition des données au sein de l'équipe de recherche, voire ultérieurement.

Exemple de réponse:

- Chaque cahier de terrain scanné est classé dans un dossier numéroté préfixé par les initiales de l'auteur type < XY_Cahier_n >.
- Chaque page numérisée est nommée d'après le schéma suivant : <initiales de l'auteur>_<année de production du cahier>_< ?? >_page_<n° de page> / <producteur_sujet_type de document_AAAAMMJJ_version.extension >
- Les entretiens retranscrits seront chacun affectés d'un numéro en fonction de l'enquête ou de la sous-enquête à laquelle il correspond et du pays et ville de collecte.

Recommandations:

Quelques outils :

- Gestion et organisation des fichiers, conventions de nommage des fichiers, gestion des versions, etc. préconisées par l'INRAE https://www6.inrae.fr/datapartage/Gerer/Nommer-et-organiser-ses-fichiers-de-donnees/Comment
- Règles de nommage des fichiers de l'Université de Lausanne : https://wp.unil.ch/gedunil/2019/07/les-regles-de-nommage/
- Outils libres de renommage des fichiers : ReNamer http://renamer.fr.softonic.com/ et Ant Renamer http://www.antp.be/software/renamer/fr
- Règles d'enregistrement des domaines Internet : https://www.afnic.fr/medias/documents/Cadre_legal/afnic-charte-de-nommage-2014-12-08.pdf

2.3. S'il y a lieu mesures de contrôle de la qualité des données

Lister les approches et outils mobilisés pour l'évaluation, le contrôle et la correction/nettoyage des problèmes de qualité des données (quantitatives), dont les problèmes spécifiques de doublons, valeurs manquantes, incomplètes ou aberrantes

Exemple de réponse:

Exemple d'outil de gestion des doublons: http://www.duplicatecleaner.com/

3. Stockage et sauvegarde pendant le processus de recherche

3.1. Modalités de stockage et de sauvegarde

Exemple de réponse:

- Les données traitées seront déposées sur la plateforme https :NOMduProjet.com dédiée à l'analyse collaborative et à la codification des donnée du projet et construites selon les modèles développés et commercialisés par la société XX.
- Les données seront stockées sur un espace créé sur le Cloud Sécurisé du CNRS selon le dispositif Espace MyCoRe.
- Les jeux de données issus du dispositif STRADIV 2 à Ivory seront déposés sur le Dataverse du Cirad et partagés avec tous les partenaires du projet.
- Le projet BFF alliant à la fois des partenaires privés et publics et étudiant plusieurs plantes, il sera nécessaire de détailler l'accès aux données selon les différents cas de figure en se référant à l'accord de consortium.
- Concernant l'étude du Sorgho au Cirad, les données GBS seront accessibles aux scientifiques du partenaire Cirad.

Recommandations:

Exemples de services de simple stockage :

La **TGIR HumaNum** propose différents outils de stockage et peut vous conseiller pour opérer le choix le plus adapté en fonction des besoins et des objectifs de votre programme de recherche : https://www.huma-num.fr/services-et-outils/stocker
Entre autres exemples d'outils proposés par HumaNum :

- **SHAREDOCS** pour stocker, mettre à jour régulièrement des fichiers, échanger de façon sécurisée des fichiers de travail au sein de son programme de recherche et opérer des traitements simples (reconnaissance de caractères, transcodage de formats audiovisuels) dans le cas de masses de données réduites.
- **Huma-Num Box** pour stocker et sauvegarder de gros volumes de données ; calcul de checksum ; scripts d'automatisation.

3.2. Modalités de sécurisation des données stockées

Préciser comment seront observées les règles de base de sécurisation des systèmes d'information, outils numériques, échanges et stockage de données et du stockage. Se reporter à la sécurisation des systèmes d'information mis à disposition par les établissements de rattachement de l'unité et des chercheurs.

Exemple de réponse:

- Les données seront traitées et gérées dans un environnement sécurisé hors réseau en utilisant la technologie du bureau virtuel

- Les fichiers de données sont diffusés aux équipes de recherche productrices par une plateforme d'accès à distance sécurisée. Un comité d'accès aux données sera mis en place

Recommandations:

Niveaux de sécurisation :

- Authentification des utilisateurs des outils numériques : les certificats numériques, les mots de passe.

- Gestion des habilitations : accès aux sites et données aux seules personnes habilitées par le responsable de traitement ou du projet de recherche ou par un comité d'accès créé à cet effet.

- Sécurisation des outils (chiffrement des ordinateurs et des smartphones) et protection des réseaux informatiques internes.

- Sécurisation des échanges entre organismes, entre unités et entre chercheurs, y compris les visioconférences (au CNRS : recours à Skype entreprise).

Outils accessibles :

- Sur la sécurisation des données personnelles, Guide de la CNIL, édition 2018 : https://www.cnil.fr/sites/default/files/atoms/files/cnil_guide_securite_personnelle.pdf

- A l'échelle du projet, voir l'offre de services numériques du CNRS : https://ods.cnrs.fr/

Attention Chaque institution peut avoir sa politique institutionnelle de protection des données : Pour PRODIG, la référente est :

Mme Gaëlle BUJAN, dpd.demande@cnrs.fr, Tél unique : 03 83 85 64 26,
https://intranet.cnrs.fr/protection_donnees/donnees/Pages/default.aspx

Autres services support : le Comité de protection des personnes (CPP) https://www.iledefrance.ars.sante.fr/comites-de-protection-des-personnes-cpp; le comité éthique ou les services juridiques de l'établissement.

4. Exigences légales et éthiques, codes de conduite

4.1. Conditions de respect des dispositions légales et de sécurisation des données personnelles

Exemple de réponse:

- Les enregistrements audio des entretiens seront conservés le temps de leur transcription et de leur pseudonymisation puis seront détruits au plus tard à la fin du projet. Une procédure d'anonymisation des retranscriptions pseudonymisées des entretiens sera mise en place : les rapports d'entretiens seront repris un à un et tous les passages contenant des informations susceptibles d'identifier la personne interrogée ou celles qu'elle mentionne seront supprimés (avec indication de la suppression). En cas de doute, la suppression sera favorisée. Chacun de ces rapports d'entretien sera validé par l'un des chercheurs ayant conduit l'enquête.

- Les données nominatives et les données d'enquêtes sont stockées dans deux systèmes d'information différents. Le traitement pour les données nominatives du panel a été enregistré au registre CIL de l'INED sous le numéro 2012-CIL-0012

Recommandations:

En cas de collecte de « données personnelles » ou données nominatives et/ou données permettant l'identification des personnes par recoupement d'information (et/ou de collecte de « données sensibles »), expliquer les mesures appliquées pour satisfaire au Règlement général sur la protection des données (RGPD) qui impose des mesures de confidentialité : recueil du consentement éclairé des participants, sécurisation des données, pseudonymisation ou anonymisation des données, etc. Se reporter au RGPD (https://www.cnil.fr/fr/reglement-europeen-protection-donnees/) ainsi qu'aux documents de synthèse mis à disposition des équipes pour préciser les droits et obligation en matière de recherche, notamment en SHS (Université de Nanterre, RGPD. *Fiches pratiques à destination des chercheurs*. URL : https://recherche.parisnanterre.fr/fiches-pratiques-sur-le-reglement-general-pour-la-protection-des-donnees-rgpd-2019-894839.kjsp; INSHS, *Les sciences humaines et sociales et la protection des données à caractère personnel dans le contexte de la science ouverte* : https://inshs.cnrs.fr/sites/institut_inshs/files/pdf/guide-rgpd_2.pdf)

Outil d'aide à l'anonymisation des données :
ARX Data Anonymization Tool

Exemples de techniques d'anonymisation (avis du G29) :

- Ajout de bruit : altérer la justesse de l'information en ajoutant de l'aléa

- Permutation : Mélanger les valeurs d'attributs au sein du jeu de données

- Généralisation : Changer la granularité des valeurs pour former des groupes

Cas particulier de données librement accessibles sur les réseaux sociaux :
Situation nouvelle. Seul exemple connu : un avis de la CNIL à propos d'une recherche sur leurs impacts pour la vie privée https://www.legifrance.gouv.fr/affichCnil.do?id=CNILTEXT000036945250

4.2. Autres questions juridiques

Exemple de réponse:

- Voir paragraphe "modalités d'exécution" dans l'accord de consortium sous l'égide du GIS biotechnologies Vertes du projet Biomasse pour le Futur.

- Les droits de propriété intellectuelle sur les données du projet et leurs dérivés restent acquis par les différentes parties ayant créé ou produit ces données. Cependant, les différents membres du consortium se sont engagés à partager les données produites entre eux.

Recommandations:

Droits de propriété intellectuelle sur les données

La propriété des données est stipulée dans l'accord de consortium dont les termes doivent être négociés en amont du projet et du PGD : si possible, donner le lien vers l'accord de consortium. A défaut, les données produites appartiennent à l'institution dont dépendent les chercheurs qui les ont collectées.
Les bases de données constituent un cas particulier : une législation spécifique s'applique, en plus de droit d'auteur, celle du droit sui generis du producteur de la base.
L'ENPC (École des Ponts ParisTech) propose un logigramme à plat, pour aider à la prise de décision en matière de propriété et diffusion de données de recherche quantitatives (type base de données) : https://espacechercheurs.enpc.fr/sites/default/files/logigramme_a_plat.pdf

4.3. Autres questions éthiques et déontologiques

Recommandations:

Cas de données d'expérimentation animale, de suivi de traitements dans des populations, de recherches pouvant avoir un impact sur l'environnement. Expliquez comment sont traitées les questions éthiques. Si besoin, citer le(s) comité(s) d'éthique ayant validé le protocole.

Outil d'aide pour les questions éthiques :
https://theodi.org/article/data-ethics-canvas/

Dans le cas de données issues de ressources génétiques ou de savoirs traditionnels associés, préciser les procédures mises en place pour respecter la législation APA (Accès et Partage des Avantages): https://www.ecologique-solidaire.gouv.fr/acces-et-partage-des-avantages-decoulant-lutilisation-des-ressources-genetiques-et-des-connaissances

Pour connaitre la législation du pays fournisseur : https://absch.cbd.int/fr/countries

5. Partage des données et conservation à long terme

5.1. Sélection des données à conserver

Différentes durées de vie et d'archivage des données sont à définir, sachant qu'il n'est pas toujours possible (et probablement pas utile) de tout conserver. Préciser les types de données qui seront détruites. La sélection peut être effectuée à partir de plusieurs critères : potentiel de réutilisation scientifique, valeur de preuve, valeur historique, etc.

Précisez quels jeux de données seront partagés sur quelle(s) durée(s). Ces durées de conservation doivent être définies et indiquées de manière transparente, ceci tout particulièrement en cas de traitement de données personnelles. Ces durées peuvent être modifiées en cours de traitement.

Attention : les jeux de données non diffusés ou non rendus librement accessibles sous licence restent soumis aux exigences d'archivage (sauf destruction).

Exemple de réponse:

- Seules les données sources ainsi que les données venant en appui des résultats publiés seront conservées pour une durée minimum de 15 ans. L'ensemble des données intermédiaires seront supprimées dans l'année suivant la fin du projet.
- Les données personnelles seront conservées au maximum pendant 5 ans après la fin du projet, ainsi que les formulaires de consentement (afin de permettre aux personnes enquêtées d'exercer leur droit d'accès, de rectification et/ou d'opposition).
- Les données personnelles seront conservées au maximum pendant 5 ans après la fin du projet, ainsi que les formulaires de consentement (afin de permettre aux personnes enquêtées d'exercer leur droit d'accès, de rectification et/ou d'opposition).
- Seules des données anonymisées générées à partir des données brutes feront l'objet d'un archivage pérenne.

5.2. Conditions d'ouverture et de partage

Pour les données jugées réutilisables, il est nécessaire de les associer à des licences qui définissent les modalités du partage et de préciser les publics potentiellement destinataires ainsi que les conditions de réutilisation. On distingue notamment le fait de rendre entièrement ouverte une réutilisation possible et une réutilisation ouverte mais avec l'obligation de tracer la chaîne d'attribution.

Si besoin, expliquer les raisons qui empêchent de partager les données ou qui imposent des restrictions au partage des données ou qui justifient la mise en place d'un embargo (motifs de nature éthique, de propriété intellectuelle, de sécurité, de protection des données personnelles).

Pour les données personnelles et sensibles ne pouvant faire l'objet d'une anonymisation complète, se reporter au dispositif proposé par l'entrepôt Be Quali, en vue de n'octroyer la possible réutilisation qu'au cas par cas, sur justificatif de recherche, et ceci au sein d'un dispositif d'explicitation et de contextualisation de la recherche : https://bequali.fr/fr/nos-pratiques/

Exemple de réponse:

- Les données qui feront l'objet d'un archivage en libre accès seront distribuées sous licence Creative Commons - Attribution - Partage dans les Mêmes Conditions 4.0 International https://creativecommons.org/licenses/by-sa/4.0/
- Seules les données sources ainsi que les données venant en appui des résultats publiés seront conservées pour une durée minimum de 15 ans. L'ensemble des données intermédiaires seront supprimées dans l'année suivant la fin du projet.
- Les bases de données seront placées sous licence ODbl. Les outils et logiciels mis au point dans le cadre du traitement et de l'analyse des données seront placés sous licence GNU GPL.
- L'ensemble des données diffusées sera placé sous licence libre Etalab ou sous licence CC-BY-NC. Les conditions de la licence seront rendues accessibles aux machines sous un format ONYX PL.

Recommandations:

Pour connaître les types de licences applicables aux données de la recherche, se reporter au site CoopIST :

https://coop-ist.cirad.fr/gerer-des-donnees/rendre-publics-ses-jeux-de-donnees/6-les-principales-licences-de-diffusion-des-jeux-de-donnees

On peut notamment utiliser les licences *Creative Commons,* créées en 2002 pour la diffusion de contenus numériques (textes, images, films). Elles permettent de combiner quatre clauses, à associer selon ses besoins, pour que les auteurs expriment les droits qu'ils veulent conserver et les droits auxquels ils renoncent quant à la réutilisation leur œuvre **:** en termes d'attribution/paternité (sigle BY), d'utilisation commerciale possible ou non (sigle C ou NC), de possibilités de modification (sigle D ou sigle ND - No Derivative Works), et enfin de condition de partage (signe SA – Share Alike).

A minima, utiliser la licence CC-by 4.0 qui impose le respect de l'attribution des données à leur(s) auteur(s).

Des outils spécifiques existent pour la protection et le partage des bases de données et pour les logiciels.

5.3. Espace de préservation (plateforme ou entrepôts)

Préciser le nom de l'entrepôt où seront déposées les données pour leur sauvegarde et/ou leur archivage. Un répertoire d'entrepôts est disponible sur : http://www.re3data.org.
L'archivage pérenne ne peut se faire que sur des plateformes spécifiques (CINES), les entrepôts de données ne possédant pas, à quelques exceptions près, cette possibilité.

Exemple de réponse:

- Les fichiers de données anonymisées seront déposés dans la plateforme DataSuds
- Les données relatives aux sciences de la terre et environnementales ont été diffusées via l'entrepôt ouvert PANGAEA et feront également

l'objet d'un "data paper" dans la revue X.

Recommandations:

Entrepôts disponibles :
- DataSuds IRD : https://data.ird.fr/connexion-a-datasuds/
- Nakala (Huma-Num) : https://www.nakala.fr/
- beQuali (CDSP – Sciences Po) : Données d'enquêtes qualitatives : https://bequali.fr/fr/
- Didómena (EHESS) : Anthropologie, Anthropologie sociale, Anthropologie historique, ethnographie et ethnologie : https://didomena.ehess.fr/
- Zenodo (CERN) : SHS, Sciences et Technologies, Vie et Santé : https://zenodo.org/
- InDoRES portail d'Inventaire des Données de la Recherche en Environnement et Sociétés (INEE -BBEES) : Biodiversité, Ecologie, Environnements, Sociétés : https://bbees.mnhn.fr/, http://www.indores.fr/

Liste des hébergeurs agréés pour les données de santé :
https://esante.gouv.fr/labels-certifications/hds/liste-des-herbergeurs-agrees

Plateformes d'archivage :
- CINES (Centre Informatique National de l'Enseignement supérieur) : Archivage des données et documents numériques https://www.cines.fr/archivage/
- Huma-Num : service d'archivage à long terme des données numériques en SHS. Pour cette activité, il est en lien avec le CINES https://www.huma-num.fr/services-et-outils/archiver

5.4. Méthodes ou outils logiciels nécessaires pour accéder et utiliser les données

Préciser tout logiciel et toute information nécessaires à la compréhension et à l'accès aux données (codes, abréviations, versions des logiciels de lecture, documents explicatifs…)

Exemple de réponse:

- Dans chaque jeu de données sous Excel, un onglet est destiné à la description des variables observées.
- JBrowse du portail BFF pour les fichiers BAM et VCF sinon Tabix ou Integrated Genome Viewer (IGV).
- La documentation associée aux données d'enquête comprend le questionnaire et le manuel des enquêteurs.

Recommandations:

Pour les logiciels (liste non exhaustive) :

- Open Software Licence (OSL): https://opensource.org/licenses/OSL-3.0
- GNU-GPL: http://www.gnu.org/licenses/gpl.html
- GNU-FDL (Free Documentation Licence) https://www.gnu.org/licenses/fdl-1.3.fr.html

5.5. Attribution d'un identifiant unique et pérenne

Il existe différents types d'identifiant numérique (DOI, Handle). Les entrepôts de données attribuent un DOI ou un Handle.

Exemple de réponse:

Le dataverse du CIRAD fournira à chaque jeu de données déposé un DOI au format suivant : doi:10.18167/DVN1/LWT7BG

6. Responsabilités et ressources en matière de gestion des données

6.1. Responsabilité de la gestion des données et de sa qualité

Recommandations:

Il est recommandé de **produire un tableau récapitulatif des personnes impliquées dans le projet et des rôles assumés** : (dans l'idéal) coordinateur scientifique, data manager assurant la qualité des données et du DMP, documentaliste chargé d'identifier et de proposer des référentiels, des ontologies et de standards de métadonnées ainsi que des entrepôts de données (institutionnels, thématiques).

Dans le cas de recherche conduite dans le cadre de partenariats impliquant plusieurs entités, relevant de tutelles différentes, associant parfois des acteurs publics et des acteurs privés, il est impératif de prévoir en amont, dans le cadre d'une convention de partenariat (type accord de consortium) quelles qualités auront les partenaires impliqués dans le projet : responsable de traitement, co-responsable de traitement ou sous-traitant. Ces contrats de collaboration doivent permettre d'identifier les rôles et les obligations de chacun, notamment en matière de sécurisation des données. Le porteur scientifique du projet est le responsable naturel du traitement, sous couvert de son directeur d'unité.

6.2. Ressources (budget et temps alloués) dédiées à la gestion des données

Exemple de réponse:

- Un ETP de [n]% a été budgété pour couvrir la gestion, la documentation et la curation des données [montant]. Des formations spécifiques en matière d'analyse et de curation de données ont été prévues en vue de la gestion du "big data" [montant].
- Coût de la maintenance de la base de données et de création d'un site web.

Recommandations:

Estimer les coûts pour rendre les données FAIR et décrire comment ces coûts seront couverts (au besoin demande de financement).
Décrire les différents types de coûts : équipements et logiciels requis (en addition à ceux existants fournis par l'institution), besoins additionnels de ressources humaines, d'expertise ou de formation, charges imposées par les entrepôts de données et les sites d'archivage à long terme.
Eléments pour estimer le coût de la gestion des données sur : https://www.ukdataservice.ac.uk/manage-data/plan/costing