CIRAD: CIRAD-TEMPLATE

1. Informations sur le projet

Version du PGD et date d'édition

Exemple de réponse:
v2, éditée le 2018-02-25

Acronyme et nom complet du projet

Exemple de réponse:
STRADIV - System approach for the TRAnsition to bio-DIVersified agroecosystems,
BFF - Biomass For the Future

Lister ici les principaux bailleur(s) / financeur(s)

Exemple de réponse:
Fondation Agropolis, ANR, Ademe, Interreg (Sudoe), etc.

Identifiant de l'appel à projet

Exemple de réponse:
AAP ACC 2017, AAP PIA2 cid56862, etc.

Thématique de l'appel

Exemple de réponse:
Changement climatique, Biotechnologies et Bioressources, etc.

Référence de la convention

Exemple de réponse:
ANR-11-BTBR-0006

Date de début du projet

Exemple de réponse:

Recommandations:
Date au format ISO : YYYY-MM-DD

Date de fin du projet

Recommandations:
Date au format ISO : YYYY-MM-DD

Nom de l'institution coordinatrice du projet

Exemple de réponse:
Cirad

Responsable scientifique du projet au Cirad

Exemple de réponse:
Nom Prénom, CIRAD, UR XXX, Montpellier, France, prenom.nom@cirad.fr
Recommandations:
Nom Prénom, institution, collectif de rattachement, ville, pays, email
+ d'informations sur l'affiliation

Objectifs du projet

Exemple de réponse:
Définir les conditions nécessaires à la transition écologique

Mots-clés

Exemple de réponse:
Biodiversité, Anatomie végétale, Télédétection, Fouille de données
Recommandations:
Privilégier quelques mots-clés Agrovoc (Thésaurus Agrovoc FAO), à compléter si nécessaire par des mots-clés spécifiques

Partenaires du projet

Exemple de réponse:

INRA (France), ISRA (Sénégal)

Recommandations:

Nom de l’institution (pays).

Mentionner uniquement les partenaires ayant participé à la collecte, à l’analyse ou au traitement des données.

2. Description du(des) jeu(x) de données

Description du jeu de données

Exemple de réponse:

Données recueillies sur le dispositif STRADIV (60*9*15m) à Ivory (Madagascar) + 17 parcelles paysannes : Itk, 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.

Recommandations:

Préciser entre autres le périmètre, l'échelle, la couverture temporelle, l’application éventuelle d’un processus qualité, etc.

Nature du jeu de données

Exemple de réponse:

- - Données expérimentales ou d’observation,
- - Images satellitaires,
- - Séquences nucléiques de sorgho de type short reads de 100 pb, digérées par Apekl,
- - Alignements nucléiques sur génome de référence du sorgho,
- - Petites variations structurales (SNP, INDEL),
- - Enquêtes,
- - Données brutes ou dérivées, collections physiques, modèles, logiciels, etc...

Pays de recueil des données

Recommandations:

Indiquez le ou les pays où les données ont été collectées

Période de recueil des données

Exemple de réponse:

- 2012-2013
- juillet-septembre 2015

Méthodes de production du jeu de données

Exemple de réponse:

- - Dispositif statistique expérimental
- - Calcul d'indice de végétation NDVI
- - HiSeq 2000 Sequencing System - Illumina de la plateforme de génomique GenoToul (GeT)
- - Interviews effectuées à Dakar selon une grille fixée

Recommandations:

Méthodes et outils utilisés pour acquérir et traiter les données

Outils de traitement

Exemple de réponse:

- - Excel, Stata, SPSS
- - Qgis,
- - Matlab v8 avec Signal Processing toolbox,
- - Tassel 5 GBS v2 Pipeline,
- - R...

Recommandations:

Logiciel(s) utilisé(s) (préciser la version si besoin), algorithme, etc.

Volumétrie prévisionnelle

- Entre 50 et 100 Mo
- Entre 1 et 10 Go
- Entre 10 et 50 Go
- Moins de 50 Mo - défaut
- Entre 100 Mo et 1 Go
- Plus de 50 Go

Exemple de réponse:

- 100 Mo
- 1 Go
- ~ 250 Go
- 5 To

Recommandations:

Estimation de la taille du jeu de données

Articles déjà publiés en lien avec ce jeu de données

Exemple de réponse:

Cédric Gaillard, Sofyan Martin, Pierre-Marie Bosc, Jalila El-Ati, Marie Claude Dop, Tarek Trabelsi, Marie-Josephe Amiot et Sandrine Dury. 2018. Explorer les liens entre agriculture et sécurité alimentaire : une enquête auprès des femmes du gouvernorat de Sidi-Bouzid en Tunisie. *Cahiers Agricultures,* 27: 15501. https://doi.org/10.1051/cagri/2018005

Recommandations:

Préciser la référence des articles, avec le lien, de préférence de type DOI

Réutilisation de données existantes

Exemple de réponse:

- - Données statistiques FAOSTAT sur la production de sorgho, libres d'utilisation
- - Données Land Cover Map 2000, Center for Ecology and Hydrology, Forest Research has licence agreement

Recommandations:

Si ce jeu de données s'appuie sur la réutilisation de données existantes, citez leur origine.
Si du matériel protégé par des droits spécifiques est utilisé au cours du projet, précisez les éventuelles contraintes d'accès et/ou réutilisation.
Les données déjà existantes que vous allez utiliser sont-elles protégées par une licence qui limite leur réutilisation ? Précisez si besoin les modalités de réutilisation et les moyens utilisés pour vous mettre en conformité avec ces modalités.
Allez-vous combiner vos nouvelles données avec ces anciens jeux de données ? Là encore vérifiez qu'une licence de réutilisation permet ce cas de figure.

Potentiel de réutilisation des données

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.

Recommandations:

Quelles seraient les réutilisations possibles du jeu de données ?
Quelles pourraient être les perspectives d'application ou de développement ?
Si il existe déjà des données similaires, quel est l'apport de ce jeu de données ?

Publics cibles

Exemple de réponse:

Communauté scientifique, secteur privé, grand public, décideurs politiques, etc.

Recommandations:

Quel public pourrait potentiellement être intéressé par le jeu de données produit ?

3. Normes et métadonnées

Description des métadonnées

Recommandations:

Quelles métadonnées décrivent ce jeu de données ? Comment ont-elles été produites (cahier de laboratoire, GPS, type d'instrument, entrée manuelle) ? Par qui ont-elles été produites ?
Ajoutez un score de qualité si disponible.

Exemple de réponse:

- - Les métadonnées sont collectées par un scientifique et s'organisent selon les champs recommandés par le standard MIAPPE : General metadata, Timing and location, Biosource, Environment, Treatments, Experimental design, Sample collection, processing, management, Observed variables.
- - Méta-données propres aux images SPOT : résolution spatiale, champ spectral, couverture temporelle.

Standards de métadonnées

Exemple de réponse:

- - MIAPPE au format ISA-Tab et Crop Ontology
- - Enquêtes selon la norme DDI

- - Standard spécifique proposé par le consortium donné dans la section précédente

Recommandations:

Citez les standards utilisés, précisez s'il s'agit du standard recommandé pour la discipline. Si aucun standard disciplinaire n'existe, expliquez la solution choisie. Pour s'informer sur les standards disciplinaires :
- https://fairsharing.org
- http://www.dcc.ac.uk/resources/metadata-standards

Format des données

Exemple de réponse:

CSV, JPEG, TXT, FASTQ, BAM, VCF…

Recommandations:

Privilégiez les formats ouverts ou largement utilisés afin de faciliter le partage
https://www.ukdataservice.ac.uk/manage-data/format/recommended-formats

Gestion des fichiers

Exemple de réponse:

Arborescence
Chaque cahier de terrain scanné est classé dans un dossier numéroté préfixé par les initiales de l'auteur ; ex. : « XY_Cahier_1".
Règles de nommage
Chaque page numérisée est nommée d'après le schéma suivant : <initiales de l'auteur>_<année de production du cahier>_< ?? >_page_<n° de page>

Recommandations:

Gestion et organisation des fichiers, conventions de nommage des fichiers, gestion des versions, etc.
https://intranet-data.cirad.fr/gerer-ses-donnees/organisation-nommage-et-formats-des-fichiers

4. Responsabilités, droits de propriété intellectuelle

Responsable de la collecte des données

Exemple de réponse:

Nom Prénom, CIRAD, UR XXX, Montpellier, France, prenom.nom@cirad.fr

Recommandations:

Nom Prénom, institution, collectif d'appartenance, ville, pays, email
+ d'informations sur les affiliations Cirad :
https://intranet-dist.cirad.fr/rediger-et-publier/affiliations

Responsable de l'analyse des données

Exemple de réponse:

Nom Prénom, CIRAD, UR XXX, Montpellier, France, prenom.nom@cirad.fr

Recommandations:

Nom Prénom, institution, collectif d'appartenance, ville, pays, email

Détenteur des droits de propriété intellectuelle

Exemple de réponse:

CIRAD

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

Recommandations:

Qui détiendra les droits de propriété intellectuelle pour les données créées lors du projet ?

La propriété des données est stipulée dans l'accord de consortium dont les termes doivent être négociés en amont du projet et du PGD : donnez le lien vers l'accord de consortium si possible.

Aspects éthiques

Recommandations:

Prendre en compte, entre autres, les éléments mentionnés dans le guide qui accompagne les projets H2020 : How to complete your ethics self-assessment (July 2016*)*
http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/ethics/h2020_hi_ethics-self-assess_en.pdf

Exemple de réponse:

- - Données personnelles (récoltées lors d'enquêtes) anonymisées / agrégées
- - Données récoltées lors d'expérimentations animales validées par un comité d'éthique
- - Recherche effectuée en partenariat avec des pays hors Europe : les recommandations du manuel H2020 ont été respectées
- - En attendant que la question soit tranchée, les données omiques issues de ressources génétiques obtenues dans le cadre du Protocole de Nagoya sont soumises aux mêmes règles notamment en ce qui concerne l'accès et le partage des avantages (APA). On s'assurera qu'avant la fin du projet les partenaires accèdent aux ressources numériques auxquelles

ils ont droit selon l'accord de consortium BFF et sachent se servir des outils d'analyse associés- (évaluer la nécessité de documentations et de formation).

5. Partage, diffusion et réutilisation des données

Pendant le projet

Recommandations:

Modalités de partage et d'accès aux données pour les partenaires du projet ou pour d'autres personnes.

Exemple de réponse:

- - 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.

Après le projet

Exemple de réponse:

- - Les jeux de données issus du dispositif STRADIV 2 seront publiés en accès libre, après une période d'embargo de 3 ans pour la publication scientifique.
- - Les données seront publiques soit via le NCBI (fastq) soit via le portail BFF (Bam et VCF).

Recommandations:

Précisez quels jeux de données seront partagés.
Modalités de diffusion et d'accès aux données : accès libre ou restreint, période d'embargo...

Type de licence

Exemple de réponse:

CC-BY, CC0, Open Knowledge Foundation Open database License, Licence ouverte...

Recommandations:

Type de licence utilisé pour la diffusion.
Pour plus d'informations, voir sur le site CoopIST : Les principales licences de diffusion des jeux de données

Entrepôt de données

Exemple de réponse:

Dataverse CIRAD, Huma-Num, Ortolang, NCBI SRA, GenBank, Pangaea...

Recommandations:

Nom de l'entrepôt où seront déposées les données, s'il est identifié
Un répertoire d'entrepôts est disponible sur : http://www.re3data.org

Identifiant numérique du jeu de données attribué par l'entrepôt

Exemple de réponse:

DOI

Recommandations:

Il existe différents types d'identifiant numérique (DOI, Handle, POI).
Le dataverse du CIRAD fournit un DOI au format suivant : doi:10.18167/DVN1/LWT7BG

Date de diffusion

Exemple de réponse:

2021-01-01

Recommandations:

Date à laquelle le jeu de données a été rendu accessible dans un entrepôt ou par publication.
Date au format ISO : YYYY-MM-DD

Lecture des données

Recommandations:

Précisez 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.

Données sensibles

Exemple de réponse:

- - Données d’enquête non anonymisées
- - SNP pouvant servir dans le cadre de kit de marqueur pour aide au phénotypage

Recommandations:

Présence de données sensibles justifiant une dispense au principe de diffusion : données protégées, personnelles, stratégiques, issues de partenariats privés, etc.

6. Archivage et conservation

Stockage et sauvegarde pendant le projet

Exemple de réponse:

- Poste de travail personnel de <nom de personne>, B2, Cirad Lavalette, Montpellier
- Disque dur externe de l’équipe xxx (UR yyy).
- Le contenu du disque dur externe est synchronisé avec un serveur Cirad tous les jeudis.
- Sur l’espace de stockage du HPC data center du plateau de bioinformatique de l’UMR AGAP

Recommandations:

Support, localisation géographique, volumétrie intermédiaire, politique de sécurité informatique

Après le projet, procédures de conservation à long terme

Exemple de réponse:

- Jeu de données déposé sur Zenodo : doi:10.167/toto21
- Les données seront sauvegardées sur l’espace de stockage possédant un backup (gs7k) du HPC data center du plateau de bioinformatique AGAP et sur l’entrepôt NCBI
- Les données climatiques, non reproductibles, seront stockées sur le Dataverse Cirad.

Recommandations:

Quelles sont les données à conserver sur le long terme ?
Support d'archivage (ex. : serveur local, serveur distant, disque dur externe, entrepôts), localisation

Durée de conservation

Exemple de réponse:

BnF : 5 ans

Recommandations:

Durée préconisée (tenir compte des exigences légales et/ou réglementaires existantes)

Volume final des données archivées

Exemple de réponse:

- 1 To
- ~ 250 Go

Recommandations:

Une fois les données à archiver choisies, reporter leur volume

Destruction de données

Exemple de réponse:

- Les noms des agriculteurs interviewés seront supprimés de l’enquête sur les pratiques culturales
- Suppression des fichiers intermédiaires de l’analyse de données

Recommandations:

Précisez si certains types de données seront détruits (ex: données à caractère personnel conformément aux recommandations de la CNIL).

Responsable de l’archivage et conservation

Recommandations:

Nom et affiliation du responsable de l’archivage et de la conservation des données. Ou organisme producteur de l'entrepôt si celui-ci garantit l'archivage pérenne

Exemple de réponse:

- - Nom Prénom, Institution, Collectif d'appartenance, Ville, Pays
- - EMBL
- - BnF

Coûts associés

Exemple de réponse:

- BnF : 1 To = 2 545 €, archivage complémentaire sur un disque : 315 € HT
- 300 € / To / an

Recommandations:

Coûts financiers, matériels et humains associés à l'archivage et à la conservation des données et moyens mis en place pour les couvrir (en particulier après la fin du projet).

Costing tool and checklist :
www.data-archive.ac.uk/media/247429/costingtool.pdf