DMP du projet "PGD de la PGTB"

Plan de gestion de données créé à l'aide de DMP OPIDoR, basé sur le modèle "INRA - Trame Structure" fourni par INRAE - Institut national de recherche pour l'agriculture l'alimentation et l'environnement.

Renseignements sur le plan

Titre du plan	DMP du projet "PGD de la PGTB"
Version	Version intermédiaire
Domaines de recherche (selon classification de l'OCDE)	
Langue	fra
Date de création	2019-07-02
Date de dernière modification	2022-11-23
Identifiant	3281

Renseignements sur le projet

Titre du projet	PGD de la PGTB
Résumé	INRAE - PGTB Site de Recherches Forêt Bois de Pierroton Plateforme de génotypage et séquençage. 69 route d'Arcachon Batiment ARTIGA 33612 Cestas Cedex - France https://pgtb.fr/ La plateforme Génome Transcriptome de Bordeaux (PGTB) est une infrastructure technologique adossé à l'UMR BIOGECO ouverte à l'ensemble de la communauté scientifique académique et aux acteurs industriels, au niveau national et international. La PGTB est sous la double tutelle INRAE et Université de Bordeaux et fait partie de l'Infrastructure de Recherche INRAE Genomics. La PGTB est par ailleurs membre du réseau France Génomique, qui regroupe les principales plateformes de génomique au niveau national. Elle est membre de la Fédération des Plateformes Labellisées de l'Université de Bordeaux et elle est labellisée Plateforme stratégique par l'INRAE (CNOC) et le GIS IBISA.
Sources de financement	• IBISA, public, privé, finaceurs récurrents + financeurs projets :

Produits de recherche :

1. Default research output (Jeu de données)

Contributeurs

Nom	Affiliation	Rôles
Ehrenmann François	INRAE	
Guichoux Erwan	Université de Bordeaux	• Coordinateur du projet • Personne contact pour les données • Responsable du plan de gestion de données

DMP du projet "PGD de la PGTB"

Informations sur la structure

Nom de la structure

Plateforme Génome Transcriptome de Bordeaux (PGTB)

Type de structure

- ISC (Infrastructure Scientifique Collective)
- Plateforme, plateau technique
- ISC (Infrastructure Scientifique Collective)
- Plateforme, plateau technique

INRAE - PGTB
Site de Recherches Forêt Bois de Pierroton
Plateforme de génotypage et séquençage.
69 route d'Arcachon Batiment ARTIGA
33612 Cestas Cedex - France
doi:10.15454/1.5572396583599417E12
https://pgtb.fr/
La plateforme Génome Transcriptome de Bordeaux (PGTB) est une infrastructure technologique adossé à l'UMR BIOGECO ouverte à l'ensemble de la communauté scientifique académique et aux acteurs industriels, au niveau national et international. La PGTB est sous la double tutelle INRAE et Université de Bordeaux et fait partie de l'Infrastructure de Recherche INRAE Genomics. La PGTB est par ailleurs membre du réseau France Génomique, qui regroupe les principales plateformes de génomique au niveau national. Elle est membre de la Fédération des Plateformes Labellisées de l'Université de Bordeaux et elle est labellisée Plateforme stratégique par l'INRAE (CNOC) et le GIS IBISA.

Identifiant de la structure
Préciser le fournisseur de l'identifiant (ISNI, VIAF, FundRef, DataCite...).

PGTB

Responsabilités dans la structure

Nom, Prénom	Courriel	Rôle
Lepais, Olivier	olivier.lepais@inrae.fr	Responsable scientifique
Delhaes, Laurence	laurence.delhaes@chu-bordeaux.fr	Responsable scientifique
Guichoux, Erwan	erwan.guichoux@inrae.fr	Responsable technique et responsable PGD

Etablissement(s) tutelle(s)

INRAE & Université de Bordeaux

Département de rattachement Inra

- EFPA
- EFPA

ECODIV

Financeur(s) (*permettant l'acquisition des jeux de données – hors projet)*

- INRAE
- Université de Bordeaux
- Région Nouvelle Aquitaine
- Etat Français
- GIS-IBiSA
- CNOC
- Recettes

Informations sur le plan de gestion

DOI (version publiée du plan de gestion)

Pas encore publié

Historique des versions

Date	n° de version	Status	Auteur	Affiliation de l'auteur *(se reporter à l'annuaire Inra)*	Validé par	Validé le

---

Présentation générale des données

Mode d'obtention des données

- Données générées par la structure
- Données produites par un tiers

- Données générées par la structure
- Données produites par un tiers

Données expérimentales obtenues par des équipements de laboratoire :

- Données produites par des machines de séquençage.
    - Données court fragment (technologie Illumina)
    - Données long fragment (technologie Oxford Nanopore)
- Données produites par des machines de génotypage.
-
    - Données de génotypage SNP et INDEL (technologie MassArray Agena Bioscience)

Données d'analyses produites par des logiciels traitant les données de sortie de machine (Analyse).

- Génomique : Assemblage, appel de variants
- Transcriptomique : Analyse d'expression différentielle
- Métagénomique : Métabarcoding

Lien vers schéma illustrant le PGD (sur site web pgtb quand le pgd sera publié)

Origine

- Analyse
- Code
- Expérimentation
- Analyse
- Code
- Expérimentation

Équipements générant les données :

- iSeq 100 (Illumina)
- 2 MiSeq (Illumina)
- NextSeq 2000 (Illumina)
- Gridion (Oxford Nanopore Technology)
- Minion Mk1C (Oxford Nanopore Technology)
- MassArray (Agena Bioscience)
- Digital Droplet PCR QX200 (Bio-Rad)

Équipements de supports (Manipulation humide) :

- Robot STAR (Hamilton)
- Dragonfly (SPL Labtech)
- Viaflo 384 (Integra Biosciences)
- LightCycler LC480 II (Roche)
- Synergy H1 (Agilent)
- Tapestation 4200 (Agilent)
- M220 (Covaris)

Logiciels traitant la données :

- Scripts développé en interne
- Pipelines développé en interne
- Logiciels/Outils provenant de la communauté scientifique (github, publication ...)
- Logiciels commerciaux (Dragen, Typer, Epi2me, WIMP, ...)

Type de données

- Dataset
- Workflow
- Dataset

- Workflow

Dataset : Fichiers de données expérimentales ou de traitements de données
Workflow : Séquence de traitement de données (Analyse SSR-Seq, Métabarcoding principalement)

Nature des données

Données expérimentales de séquençage ou génotypage

- Séquence nucléotidique (ADN/ARN)
- Données de génotypage (SSR, SNP, INDEL)

Données dérivées ou compilées, issues de traitement et de l'analyse des données brutes

- Contrôle qualité (QC, Image, Témoins)
- Analyse de données -omique (génomique, transcriptomique, métagénomique)
- Rapport des données et des analyses
- Système d'information (Redmine, lien interne http://147.100.175.60/)

Format des données

Formats propriétaires et non-propriétaires
Multiples formats de données, dépendant de l'origine des données

- Séquençage : fasta, fastq, md5sum, fast5, zip, html
- Génotypage : xml, xlsx, png
- Analyse bioinformatique : sam, bam, fasta, fastq, vcf, gff, bed, txt, csv, html, xls, Rdata
- Programmation : Python, Bash, Awk, Perl, R, Snakemake, Yaml, Git, Markdown
- Texte : pdf, odf, doc, docx, txt
- Dataset : xlsx, xls, txt, csv
- Rapports : pdf, doc, docx, ppt, xls, xlsx, html, tar.gz, png

---

Périmètre thématique des données

- Animal Breeding and Animal Products
- Animal Health and Pathology
- Biodiversity and Ecology
- Fishes and Aquaculture
- Food Safety and Toxicology
- Forests and Forest Products
- Human Health and Pathology
- Human Nutrition and food security
- Insects and Entomology
- Microorganisms
- Omics
- Plant Breeding and Plant Products
- Plant Health and Pathology
- Soils and soil sciences
- Animal Breeding and Animal Products
- Animal Health and Pathology
- Biodiversity and Ecology
- Fishes and Aquaculture
- Food Safety and Toxicology
- Forests and Forest Products
- Human Health and Pathology
- Human Nutrition and food security
- Insects and Entomology
- Microorganisms

- Omics
- Plant Breeding and Plant Products
- Plant Health and Pathology
- Soils and soil sciences

Droits de propriété intellectuelle

Qui détiendra les droits sur les données et les autres informations créées ?

- Projet en Prestation : la PGTB est propriétaire des méthodes et l'utilisateur est propriétaire des résultats. La PGTB peut donc librement communiquer sur les méthodes, mais ne pourra communiquer sur les résultats qu'après accord écrit de l'utilisateur. En terme de valorisation, l'utilisateur doit systématiquement mentionner la PGTB dans les remerciements de" toutes les valorisations scientifiques, en cas d'une implication significative de la PGTB, le co-autorat sera de mise.
- Projet en R&D/Collaboration : l'utilisateur et la PGTB définissent la propriété intellectuelle des méthodes et des résultats d'un commun accord dans le cahier des charges. Il en sera de même dans le cas de valorisation scientifique.

Confidentialité

Identification des jeux de données contenant des données confidentielles

De manière générale, les données sont la propriété des utilisateurs et sont considérées comme des informations confidentielles. Les données resteront confidentielles jusqu'à publication par l'auteur.
Dans le cas spécifiques des projets internes, les données ont vocations a être rendues publiques (publications, poster, data.gouv).
La liste des clients et données personnelles sont listés en interne à la structure.

Quelles sont les mesures prises et les normes auxquelles il est nécessaire de se conformer pour garantir cette confidentialité ?

Anonymisation des noms de projets et des échantillons proposés aux utilisateurs.
Transfert des données uniquement à l'utilisateur.
Accès aux données limité à l'utilisateur et collaborateur

Le cas échéant, comment la confidentialité de données fournies par des personnes sera garantie lorsque les données seront partagées ou rendues disponibles pour une analyse de second niveau ?

Pas concerné
Un accès aux données à un tiers sera possible si l'utilisateur porteur du projet autorise le partage de ses données pour analyse.

Partage des données

Y a t'il une obligation de partage (ou à l'inverse une interdiction ou une restriction) ?

La donnée appartient à l'utilisateur, donc le choix lui revient.
Les données générées dans le cadre d'un projet interne financé en partie ou en totalité pour un financeur publique suit la politique d'ouverture des données selon le principe "ouvert autant que possible, fermé autant que nécessaire".

Quelles sont les réutilisations potentielles de ces données ?

L'utilisateur est libre de faire ce qu'il veux des données produites (analyse, publication...)

Les données de sortie de séquençage sont des données brutes intéropérables. En ce qui concerne le traitements de données, les versions de logiciels, les machines utilisées et le cheminement des analyses sont tracés sur Redmine. Donc le travail est reproductible. De plus la majeure partie des prestations possèdent un mode opératoire décrivant le cheminement de la prestation ce qui permet de rendre réutilisable les différentes prestations effectuées

---

La lecture des données nécessite-t-elle le recours à un logiciel ou un outil spécifique ? Si oui, lequel ?

Les données fournies à l'utilisateur sont dans un format interprétable/ intéropérable cependant différents logiciels peuvent par la suite être utilisés par l'utilisateur afin de traiter la donnée.
Cas particulier des données brutes de génotypage MassArray : obligation pour l'utilisateur d'avoir une licence Typer s'il souhaite analyser les données brutes fournies au format xml (cas très ponctuels).

Comment les données seront-elles partagées ?

Les données peuvent être partagées à l'utilisateur par différents moyens :

- Disponibles sur serveur NAS interne (https://nas-pgtp.pierroton.inra.fr:9444/index.cgi et https://biogeco-bkup.synology.me:9444/ , nécessite des droits d'accès)
- Envoi des données par Filesender. (https://filesender.renater.fr/)
- Envoi de support physique avec données (clé USB/Disque Dur).
- Utilisation de la solution de stockage Illumina Basespace

Les données pourront à terme être publiées pour la communauté scientifique (au choix de l'utilisateur) par différents moyens :

- recherche.data.gouv
- Zenodo
- NCBI

Avec qui ?

- Autre
- Autre

Les données sont uniquement partagées avec l'utilisateur, ainsi c'est à l'utilisateur de choisir avec qui il compte partager ses données produites par la PGTB.

Sous quelle licence ?

Non concerné

Organisation et documentation des données

Quels méthodes et outils sont utilisés pour acquérir et traiter les données, depuis leur acquisition jusqu'à leur mise à disposition, leur archivage ou leur destruction ?
Utiliser éventuellement un lien vers un schéma illustrant les processus

- Définition des besoins et des méthodes avec l'utilisateur
- Réception des échantillons
- Validation des échantillons (respect des conditions d'acceptabilité : conditions générales, prérequis)
- Stockage des échantillons
- Préparation des échantillons (dosage, dilution, amplification ...)
- Séquençage ou génotypage des échantillons
- Stockage et archivage des données sur le NAS
- Contrôle qualité
- Analyse bio-informatique (si dans la prestation)
- Stockage et archivage des résultats/données traités sur le NAS.
- Traçabilité des analyses réalisé (cheminement) et des métadonnées pour chaque projet sur PGTBsi-Redmine (appareil utilisé, dates, échantillons, manipulateur, réactifs, suivi d'utilisation ...). Pour chaque appareil ou traitement de données, des tableaux ou documents de suivi exhaustif sont rédigés (date, opérateur, lots de réactifs, contrôles qualité, ...).
- Rédaction rapport pour les prestations de bio-informatiques en séquençage, ou envoi d'un mail.
- Envoie des données, des résultats issus du traitement de données et rapport d'analyse à l'utilisateur.
- Destruction des échantillons et des données 2 mois après l'envoi des derniers résultats (sauf projets R&D et internes)
- Possible interaction avec l'utilisateur (réunion, rédaction échange de mail) si besoin d'informations supplémentaires ou pour valorisation scientifique.

Lien vers schéma illustrant le PGD

Quelles métadonnées seront utilisées pour accompagner le jeu de données ? Quels seront les standards, vocabulaires, taxonomies... utilisés pour décrire et représenter les données et éléments de métadonnées ? Comment les métadonnées seront-elles produites et mises à jour ?

Quel que soit le projet, l'ensemble des métadonnées sont écrites et conservées sur redmine. Chaque étape de l'analyse possède des métadonnées bien spécifique à l'étape en question. De plus les machines peuvent créer leur propres fichiers de métadonnées qui peut être conservées sur le NAS ou sur Redmine.

Métadonnées	**Origine, mode de production des métadonnées (ex : saisie manuelle, annotation atomatique...)**	**Standard, Vocabulaires associés**	**Conditions ou fréquence de la mise à jour (si applicable)** *(ex : changement de l'accessibilité*
metadonnées séquençage et génotypage	Saisie via l'interface web redmine	Vocabulaire controlé pour différentes métadonnées et vocabulaire machine.	

Une documentation complémentaire aux métadonnées est-elle nécessaire pour décrire les données et assurer leur réutilisabilité sur le long terme ?

Description de l'expérience, des manipulations, des protocoles (expérimentaux et bio-informatiques). Une documentation des métadonnées essentielles est rendue sous forme de rapport ou de mail à l'issue de la prestation en bio-informatique. De plus les interactions restent possible avec la plateforme une fois la prestation terminée.

Comment les fichiers de données sont-ils gérés et organisés : contrôle des versions, conventions de nommage des fichiers, organisation des fichiers

Les fichiers de données sont conservées sur un NAS et les métadonnées ainsi que le cheminement des analyses sont conservées sur un serveur redmine.

- Sur le NAS, les prestations peuvent être rangées par une ou plusieurs de ces catégories représentées par des dossiers et sous dossiers :

- Technologies (Illumina, Oxford Nanopore Technology, Agena Bioscience ...)
- Machines (NextSeq2000, Miseq ...)
- Activités (SSR-Seq, Métabarcoding, ...)
- Années

Puis dans le sous dossier le plus approprié à la prestation un dossier portant l'acronyme du projet sera créé qui contiendra l'ensemble des fichiers associées à la prestation. Chacun de ces dossiers contient les Inputs et Outputs des différents étapes réalisées dans le cadre de la prestation.

- Sur redmine un projet est créé pour chaque prestation, ce projet porte l'acronyme du projet. Puis chaque sous taches associées à la prestation sont associées au projet, afin que toutes les métadonnées soient reliées les unes aux autres.

D'autre part, concernant le nommage des fichiers, les machines peuvent générer leur propres fichiers suivant leurs propres conventions de nommage. En ce qui concerne les fichiers créés/générés par le personnel de la plateforme, il n'y a pas de convention de nommage clairement établie, mais le personnel veille à ce que le nom du fichier soit parlant.

Enfin, concernant les versions un tableau des suivis des versions est disponible en interne sur le NAS qui est mis à jour lorsqu'une version de machine ou de logiciel change au sein de la plateforme.

Quelle est le processus de contrôle qualité des données ?

- Contrôles pour valider la conformité et la qualité des échantillons entrants (ADN/ARN) -> précisé pour chaque analyse dans des prérequis spécifiques accessibles sur notre site web (https://pgtb.fr/documents-et-liens/).
- Contrôles internes qui valident la qualité du séquençage et du génotypage (témoin négatif, témoin positif). Dans le cas du séquençage des outils de QC sont aussi utilisées (fastQC, PycoQC)
- Possibilité de contrôle qualité dans certaines analyses bio-informatiques.

Ces contrôles sont compilés et analysés dans le cadre des certifications ISO 9001 et NFX 50-900 sous forme de KPI et d'indicateurs.

Stockage et sécurité des données

Quels sont les types de flux empruntés par les données et les supports utilisés pour les stocker ?
(Faire éventuellement un lien vers un schéma)

La PGTB utilise un NAS pour stocker ces données.

Lorsque le séquençage ou le génotypage est terminé un copier coller des données depuis la machine vers le NAS est fait.

Les espaces de stockage de ces NAS possèdent des sauvegardes à l'heure (snapshot) pour la journée en cours et à la journée pour le mois en cours. De plus il y a un archivage des données grâce à 2 copies à 2 endroits différents fait automatiquement mensuellement.

Pour les données issues du Nextseq 2000, en plus d'être copier coller vers un NAS, les données sont aussi transférées au cloud Illumina Basespace (flux/transfert intégré à la machine).

Lien vers schéma illustrant le PGD

Quelle est la volumétrie actuelle et prévisionnelle ?

La volumétrie actuelle est de 43.3 TB sur l'espace PGTP. La volumétrie disponible est de 19.54 TB.
La volumétrie actuelle est de 46.39 TB sur l'espace NextSeqSpace/ONT_backup. La volumétrie disponible est de 37.4 TB.

En ce qui concerne la volumétrie prévisionnelle, il n'y a pas de changements à prévoir avec les équipements présent sur la PGTB, car la plateforme n'a pas vocation à faire du stockage de données. Les données des utilisateurs sont censées être supprimées 2 mois après la fin du projet.
Un changement de volumétrie serait à prévoir si la PGTB acquerrait de nouveaux équipements nécessitant des plus gros espaces de stockage (car des plus gros volumes de données produites par RUN).

L'entité hébergeant physiquement les données a-t-elle une politique de sécurité pour son système d'information ?
politique locale, charte des infrastructures de recherche...

Repose sur les SI de l'unité (UMR 1202 BIOGECO), politique de sécurité en vigueur :

- dans le portail Bordeaux Aquitaine
- Portail des données INRAE
- Charte des infrastructures de recherche à l'Inra

Cela implique une sauvegarde du NAS avec réplication

Sécurité - Confidentialité : les données font-elles l'objet d'échange ou de partage avec de tiers acteurs et selon quelles modalités ? comment sont déterminés les droits d'accès aux données avant leur publication ?

Seuls les coordinateurs ou collaborateurs du projet et le personnel de la plateforme ont accès aux données tant qu'elles ne sont pas publiques.
Pas d'échanges avec des acteurs tiers par défaut, sauf dans le cas des données NextSeq 2000 stockées sur le cloud BaseSpace d'Illumina qui peuvent être accessibles au support technique lorsqu'un run est problématique (accès sécurisé et individuel).

Sécurité - Intégrité - Tracabilité : Quelles sont les mesures de protection mises en œuvre pour suivre la production et l'analyse des données ?

Sécurité

- Protection contre les virus et les intrusions géré par le site de Cestas Pierroton
- Restrictions sur le droit d'accès
- Sauvegarde des données : snapshots plusieurs fois par jour (toutes les heures) conservé sur 24h. Sauvegarde journalière. Copie/Archive sur serveur secondaire une fois par mois

---

Archivage et conservation des données

Quelles sont les données à conserver sur le moyen ou le long terme et quelles sont les données à détruire ?

La règle générale pour les prestations est que 2 mois après la clôture du projet (dernières données envoyées), les données ne sont pas conservées, seules les métadonnées le sont. Possibilité de déroger à cette règle si accord préalable avec l'utilisateur.
Pour les projets internes et les projets R&D, les données sont conservées jusqu'à leur publication selon les principes FAIR.
A plus long termes toutes les données sont détruites, la plateforme n'a pas vocation à faire de l'archivage de données. La conservation des données appartient à l'utilisateur.

Sur quelle plateforme d'archivage pérenne seront archivées les données à conserver sur le long terme ? Sinon, quelles procédures seront mises en place pour la conservation à long terme ?

Pas d'archivage sur du long terme.

Quelle est la durée de conservation des données ?

La règle générale pour les prestations est que 2 mois après la clôture du projet (dernières données envoyées), les données ne sont pas conservées, seules les métadonnées le sont.

Quelles garanties de financements couvriront les coûts associés à la conservation à long terme ?

La PGTB n'a pas vocation à conserver les données produites sur du long terme.