PGD UMR BIOGECO

Plan de gestion de données créé à l'aide de DMP OPIDoR

Créateurs du PGD : François Ehrenmann, Philippe CHAUMEIL, christophe Plomion

Affiliation du créateur principal : INRAE - Institut national de recherche pour l'agriculture l'alimentation et l'environnement

Modèle du PGD : INRA - Trame Structure

Dernière modification du PGD : 02/09/2021

Résumé du projet :

Plan de gestion de données de l'UMR Biogeco

Le programme de recherches de l'UMR est orienté vers l'analyse des mécanismes régissant l'évolution de la diversité biologique à différents niveaux hiérarchiques (communautés, espèces, populations, gènes) dans une perspective de gestion durable des ressources et des milieux. Les milieux concernés sont surtout les forêts, prairies et zones humides.

Les espèces les plus étudiées par les chercheurs de BIOGECO sont les arbres, les champignons pathogènes, les insectes et les plantes herbacées. Les recherches développées au sein de l'Unité ont pour vocation de promouvoir une analyse plus intégrée de la diversité biologique, en considérant les interactions entre espèces, populations et individus comme moteurs de son évolution. Elles s'appuient sur des dispositifs expérimentaux importants et bénéficient du soutien de deux infrastructures scientifiques collectives spécialisées en génotypage-séquençage et sur l'analyse des propriétés du bois, ainsi que de celui de plusieurs unités expérimentales Inra (notamment l'Unité expérimentale Forêt Pierroton). L'UMR produit par conséquent de grandes quantités de données, de natures très variées.

Chercheur Principal : Christophe Plomion

Identifiant ORCID : https://orcid.org/0000-0002-3176-2767

Contact pour les Données : François Ehrenmann

PGD UMR BIOGECO

Informations sur la structure

Nom de la structure

UMR1202 BIOGECO Biodiversité, Gènes et Communautés

Type de structure

- Unité de recherche, Unité ou Installation Expérimentale

INRA - UMR1202 BIOGECO
Site de Recherches Forêt Bois de Pierroton
69 route d'Arcachon
33612 CESTAS Cedex - FRANCE
Site web de l'unité : https://www6.bordeaux-aquitaine.inrae.fr/biogeco

L'unité UMR1202 BioGeCo, constituée d'environ 110 personnes, est basée sur le centre INRAE de la région Nouvelle Aquitaine à Bordeaux. Le programme de recherche de l'unité Mixte de Recherches BIOGECO est orienté vers l'analyse de la structure, de la fonction et de l'évolution de la biodiversité à différentes échelles du vivant (des gènes aux communautés d'organismes) dans une perspective de gestion durable des ressources et des milieux. Les recherches sont développées au sein de 7 équipes thématiques appuyés par 4 pôles de compétences métiers qui ont pour ambition commune de promouvoir une analyse intégrée de la diversité biologique, en considérant les interactions entre espèces, populations et individus comme moteurs de son évolution.

L'UMR publie des articles de rang A dans des disciplines variées avec un fort tropisme dans les domaines de l'Ecologie, Science des plantes, Foresterie, Biologie Evolutive, Sciences Environnementales, Biodiversité et Conservation, Génétique et Hérédité.

Ce document explicite la manière dont sont obtenues, documentées, analysées, disséminées et archivées les données produites par l'UMR. Ce document est conçu comme un outil pour gérer les données en intégrant la notion de cycle de vie. Le PGD s'étend donc de la production (ou la collecte) des données à leur diffusion et/ou leur archivage, en passant par leur stockage, leur traitement/curation, leur analyse et leur description. Ce document a vocation à être référencé dans les PGD de projets élaborés dans le cadre des différentes réponses à appel d'offre.

Nature des données

- observations (données de terrain uniques, non reproductibles),
- expérimentales (obtenues à partir d'appareils de mesures ou en laboratoire, peuvent être reproductibles),
- simulation (générées par des modèles informatiques ou de simulation, reproductibles)

Origine des données

- données génétiques : Génomes entiers, Marqueurs génétiques (SNP,) et épigénétiques (Méthylation), Cartes génétiques)
- données phénotypiques (Croissance, Phénologie, Architecture, Bois, Biodiversité, Transcriptome et Métabolome, Herbivorie, Prédation, Biomasse, Germination, etc...)
- données environnementales (Sol, climat)
- données informatiques (Programme, Scripts, BDD, Analyses / Statistiques, Simulation / Modélisation)
- autres types de données (Photos, Vidéos, GPS)
- associations Génotype-Phénotype, Génotype-Environnement, Phénotype-Environnement

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

1202

Responsabilités dans la structure

Nom, Prénom	Courriel	Rôle
Christophe Plomion	christophe.plomion@inrae.fr	Directeur d'unité / Directeur de recherches
Cécile Robin	cecile.robin@inrae.fr	Directrice adjointe / Directeur de recherches
Virgil Fievet	virgil.fievet@u-bordeaux.fr	Directeur adjoint / Enseignant-chercheur
François Ehrenmann	francois.ehrenmann@inrae.fr	Responsable PGD / Gestion des données
Philippe Chaumeil	philippe.chaumeil@inrae.fr	Responsable PGD / Infrastructures informatiques

Etablissement(s) tutelle(s)

- INRAE
- Université de Bordeaux

Département de rattachement Inra

- EFPA

Nouvellement ECODIV

- Site web
- Intranet

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

- Budget SE (Subvention d'Etat)
- Financement sur projet (contrats)
- Recettes

Informations sur le plan de gestion

DOI (version publiée du plan de gestion)

https://doi.org/10.15454/XS1RPM

Historique des versions

Date	n° de version	Status	Auteur	Affiliation de l'auteur *(se reporter à l'annuaire Inra)*
06/09/2019	1	Non validé	• François Ehrenmann • Philippe Chaumeil	https://intranet.annuaire.inrae.fr/afficherPersonne.action?matricule=129225M&personneRecherche.nom=Ehrenmann&personneRecherche.approchee=false&personneRecherche https://intranet.annuaire.inrae.fr/afficherPersonne.action?matricule=97620J&rec=true&personneRecherche.nom=Chaumeil&personneRecherche.approchee=false&personneRech
26/09/2019	1	Non validé	• François Ehrenmann • Philippe Chaumeil	https://intranet.annuaire.inrae.fr/afficherPersonne.action?matricule=129225M&personneRecherche.nom=Ehrenmann&personneRecherche.approchee=false&personneRecherche https://intranet.annuaire.inrae.fr/afficherPersonne.action?matricule=97620J&rec=true&personneRecherche.nom=Chaumeil&personneRecherche.approchee=false&personneRech
19/11/2019	1	Non validé	• François Ehrenmann • Philippe Chaumeil • Christophe Plomion	https://intranet.annuaire.inrae.fr/afficherPersonne.action?matricule=129225M&personneRecherche.nom=Ehrenmann&personneRecherche.approchee=false&personneRecherche https://intranet.annuaire.inrae.fr/afficherPersonne.action?matricule=97620J&rec=true&personneRecherche.nom=Chaumeil&personneRecherche.approchee=false&personneRech https://intranet.annuaire.inrae.fr/afficherPersonne.action?matricule=10992Y&personneRecherche.nom=Plomion&personneRecherche.approchee=false&personneRecherche.prer
12/03/2020	2	Non validé	• François Ehrenmann • Philippe Chaumeil • Christophe Plomion	https://intranet.annuaire.inrae.fr/afficherPersonne.action?matricule=129225M&personneRecherche.nom=Ehrenmann&personneRecherche.approchee=false&personneRecherche https://intranet.annuaire.inrae.fr/afficherPersonne.action?matricule=97620J&rec=true&personneRecherche.nom=Chaumeil&personneRecherche.approchee=false&personneRech https://intranet.annuaire.inrae.fr/afficherPersonne.action?matricule=10992Y&personneRecherche.nom=Plomion&personneRecherche.approchee=false&personneRecherche.prer
28/05/2021	3	Validé	• François Ehrenmann • Philippe Chaumeil • Christophe Plomion	https://intranet.annuaire.inrae.fr/afficherPersonne.action?matricule=129225M&personneRecherche.nom=Ehrenmann&personneRecherche.approchee=false&personneRecherche https://intranet.annuaire.inrae.fr/afficherPersonne.action?matricule=97620J&rec=true&personneRecherche.nom=Chaumeil&personneRecherche.approchee=false&personneRech https://intranet.annuaire.inrae.fr/afficherPersonne.action?
02/09/2021	4	Validé	• François Ehrenmann • Philippe Chaumeil	https://intranet.annuaire.inrae.fr/afficherPersonne.action?matricule=129225M&personneRecherche.nom=Ehrenmann&personneRecherche.approchee=false&personneRecherche https://intranet.annuaire.inrae.fr/afficherPersonne.action?matricule=97620J&rec=true&personneRecherche.nom=Chaumeil&personneRecherche.approchee=false&personneRech

Présentation générale des données

Mode d'obtention des données

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

- Données de terrain capturées en temps réel; manuellement ou automatiquement; généralement uniques et irremplaçables
  - Données d'observation en forêt
    - relevés sur arbres milieu naturel, expérimentation (croisements contrôlés, plantations comparatives, parcs à clones, ...), sur différents tissus (feuilles, rameaux, bourgeons, racines, écorces,..)
    - relevés et prélèvements en entomologie
    - relevés pédoclimatiques
    - imagerie
  - Données d'observation sur échantillons (végétal, animal, champignons, ...)
- Données expérimentales capturées grâce à des équipements de labos; souvent reproductibles; peuvent être très coûteuses
  - Données obtenues à partir d'ADN ou d'ARN d'échantillons préparés par la structure et analysés par le PCM PGTB
- Données computationnelles, de modèles ou de simulations; Générées depuis des modèles d'essais ; reproductibles (s'il existe une documentation détaillée et accessible)
  - Modèles climatiques
  - Modèles de simulation d'évolution génétique et ou démographique
- Données dérivées ou compilées; Issues du traitement et de l'analyse des données brutes ; reproductibles; très coûteuses
  - Bases de données compilées
  - text/data mining
  - modèles 3D
- Données de référence ou canoniques; Préalablement organisées, gérées voire publiées
  - Banques de données de séquences ADN
  - Portails de données

- Codes informatiques / Logiciels (Systèmes d'information, applications web, scripts, logiciels, ...)
- Photos / images

Cycle de vie des données UMR Biogeco
Interface web "Logigramme origine et flux de données"

Ce cycle et cette interface décrivent les différents processus de gestion des données:

- collecte et gestion des échantillons
- traitement et curation des données
- description des données
- stockage sur le poste de travail
- transfert des fichiers vers un espace de stockage sécurisé (NAS de centre)
- mise en base de données dans certains cas
- partage à la demande vers d'autres Systèmes d'Information
- droits et propriété
- archivage

Origine

- Analyse
- Code
- Expérimentation
- Observation
- Simulation, modélisation
- Autre : à préciser dans la zone "Informations supplémentaires"

Science participative

Type de données

- Collection
- Dataset
- Image
- Model
- Software

- Données de terrain et d'observation
  - la collecte des données en milieu naturel ou en laboratoire peut s'effectuer sur papier, saisie électronique de terrain, ou automatiquement avec des instruments de type hobos, capteurs, LiDAR, etc...
  - Gestion des échantillons via application web / bases de données pour identification, responsables, lieux de stockage et quantités stockées
- Données expérimentales capturées grâce à des équipements de labos
  - données de séquençage d'ADN
  - données de génotypage
  - Transcriptome
- Données dérivées ou compilées; Issues du traitement et de l'analyse des données brutes
  - Bases de données
  - Systèmes d'information

Nature des données

- observations (données de terrain uniques, non reproductibles),
- expérimentales (obtenues à partir d'appareils de mesures ou en laboratoire, peuvent être reproductibles),
- simulation (générées par des modèles informatiques ou de simulation, reproductibles)

Format des données

Spécifie le ou les formats des fichiers qui seront employés pour le versement, la distribution et éventuellement la préservation des données produites ainsi que de leurs produits dérivés.
Voir Formats de fichiers par catégorie de données

Multiples formats de données, dépendant de l'origine des données :

- Dataset : csv, xls, xlsx, json, odf, R
- Texte : pdf, html, odf, doc, docx
- Image : png, jpeg, tiff, pdf, Encapsulated PostScript EPS
- Formats génériques : rar, zip, xml
- Shapefiles : shx, shp
- Bases de données : PostgreSQL, MySQL, MariaDB, MongoDB
- Programmation : Python, Perl, Bash, R, Javascript, PHP, Ruby, html, C, C++
- Bioinformatique : fasta, fastq, bam, gff, bed, sam

Périmètre thématique des données

- Biodiversity and Ecology
- Forests and Forest Products
- Insects and Entomology
- 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 ?

L'UMR est en accord avec les principes de l'Open Data et respecte la Charte des Infrastructures de Recherche de l'INRAE. Les données sont propriété de l'INRAE. Droits spécifiques sur les données en cas de partenariat avec le privé.

Chaque article issus des travaux de l'unité doit être publié en libre accès : (i) via la voie diamant (gratuité totale), (ii) la voie verte (1. auto-archive d'un preprint non évalué par les pairs, par exemple dans BIORXIV, 2. archive d'un preprint recommandé par exemple par le modèle des PCI, ou 3. d'un post-print non formaté accepté dans une revue payante, modulo un embargo de 6 mois, par exemple dans HAL) ou (iii) la voie dorée (via le paiement d'APC le plus souvent).

Le schéma suivant illustre les interactions possibles entre voies vertes et dorées:

https://www.openaire.eu/openaire-h2020-fact-sheet-for-researchers-2017/view-document

Confidentialité

Identification des jeux de données contenant des données confidentielles

Un SI local permettra de décrire l'ensemble des jeux de données (via leurs métadonnées) produits au sein de l'UMR. Un champ spécifique sera dédié afin de définir si tout ou partie des données sont confidentielles. Cette confidentialité dépend fortement de la nature des données (essentiellement dans le cas des données géographiques soumises au RGPD) ainsi que des partenaires identifiés.

Interface web de saisie de métadonnées pour un jeu de données

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

- Au niveau technique, attribution de droits spécifiques d'accès aux répertoires et aux fichiers (sur stockage réseau NAS) aux différents utilisateurs en fonction de la confidentialité indiquée dans le SI.
- En interne, les utilisateurs pourront accéder au SI de manière simple (NAS dans espace sécurisé dédié à l'UMR) et de l'extérieur, ils devront obligatoirement s'authentifier via VPN pour y accéder.
- Les accès aux SI reposant sur des bases de données utilisent la gestion de droits type ACL.

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 ?

Les données ayant vocation à être mises à disposition sur les différents SI ou entrepôts seront accessibles uniquement grâce à la demande de création d'un compte utilisateur permettant l'accès à ces données. Le propriétaire des données reste seul responsable de la validation de cette demande de création de compte. Si la demande de création de compte est validée, l'utilisateur doit s'engager à respecter une charte d'utilisation des données (charte utilisateur en cours de rédaction).

Partage des données

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

L'UMR est soumise aux règles institutionnelles (formalisées par la charte pour le libre accès aux données et aux publications). Seules les données acquises dans le cadre d'un partenariat privé peuvent nécessiter une licence particulière.
Ouverture des données : ... aussi ouvert que possible ... aussi fermé que nécessaire
https://datapartage.inrae.fr/content/download/3818/40488/version/1/file/LogigrammePPMD-version+8+janvier+2021.pdf

Quelles sont les réutilisations potentielles de ces données ?

- Réutilisation pour nouvelles études ou études complémentaires
- Aggrégation avec de nouvelles données (méta-analyses)
- Modélisation ou simulation

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

A définir selon la nature et l'origine des données et selon les types de logiciels utilisés. Les formats non propriétaires sont fortement recommandés. *Voir Formats de fichiers par catégorie de données*.
La liste ci-dessous est non exhaustive.

- XLS ou XLSX ou CSV : Excel ou LibreOffice ou OpenOffice
- DOC ou DOCX : Word ou LibreOffice ou OpenOffice
- TXT, DAT, HTML : notepad, geany, gedit, atom, etc...
- Bases de données : client postgres / ligne de commande / pgadmin / phpgadmin / phpMyadmin
- Fichiers R : Rstudio ou ligne de commande
- PDF : adobe reader ou evince
- PNG, JPEG, TIFF, PDF, Encapsulated PostScript EPS : Gimp, PhotoFiltre, Adobe Photoshop Elements
- RAR, ZIP, XML : WinZip, WinRAR, RAR, 7-Zip
- SHP, SHX : QGis
- Langages PHP, Javascript, R code, Bash, Python, C, C++, PERL : éditeurs geany, atom, vim, Notepad++, gedit, GNU Emacs, Brackets

Comment les données seront-elles partagées ?

- **Interne** : Prise en charge intégrale ou partielle par l'UMR en utilisant les services existants
  - stockage réseau (NAS UMR) des jeux de données et métadonnées via l'interface web dédiée adossée au PGD de structure
    - NAS Pierroton : https://nas-pgtp.pierroton.inra.fr/ et NAS Fac Talence : https://biogeco-u.synology.me:5001/
      - Espace NAS pour des jeux de données et métadonnées via l'interface web dédiée adossée au PGD de structure : File station -> PGD
      - Espace NAS pour les Stagiaires / Thésards : File station -> STAGES_THESES
      - Espace NAS pour les projets : File station -> Projets
  - Bases de données (Quercus Portal, Pinus Portal, Populus Map, R-SYST)
  - Systèmes d'information de l'UMR (site web, OAK GENOME, TREEPEACE)
  - Epaces sharepoint (personnel onedrive ou collectif de projet)
- **Interne** / **Externe**
  - Dissémination prise en charge par le projet à travers des interfaces tel qu'un site web (par exemple : www.oakgenome.fr). Si cette solution est choisie, il est recommandé que le ou les producteurs de données fassent le nécessaire afin d'effectuer l'archivage éventuel de celles-ci après que la période de dissémination des données se termine.
  - Entrepôts institutionnels gérés par un organisme public (portail Data INRAE). Possibilité d'affecter des droits par groupes ou utilisateurs.
  - Entrepôt disciplinaire (ex : Observatoire Virtuel, Entrepôt de données de génomique). Possibilité d'affecter des droits par groupes ou utilisateurs.
  - Préservation des données avec une dissémination différée (période d'embargo). Dans ce cas de figure, le producteur des données passe un accord avec

un dépôt publique de données pour l'archivage des données avec une dissémination qui démarre à une date ultérieure.

- **Externe**
  - Entrepôts institutionnels gérés par un organisme public (portail Data INRAE)
    - Dataverse Biogeco
  - Partage externe automatisé: procédure d'export vers le SI national (Exemple: GnpIS)
  - Bases de données ouvertes
  - Systèmes d'information
  - Forges logicielles

Le logigramme des flux de données (Valorisation des données) recense des entrepôts de données pour le partage des données (liste non exhaustive).

Systèmes d'information et bases de données de l'UMR

Avec qui ?

- Tous (open acces)

Le Contributeur devra identifier si ses données contiennent des droits de propriété intellectuelle de tiers.
Ce peut être notamment le cas pour les données acquises par transfert, téléchargement, achat, lors d'une collaboration et, dans ce cas, même si c'est le contributeur qui a collecté les données (le contrat de partenariat pouvant définir un partage de la propriété intellectuelle sur les données).

Sous quelle licence ?

- Licence ouverte https://www.etalab.gouv.fr/licence-ouverte-open-licence (compatible CC-BY)
- Préciser comment les données peuvent être réutilisées (licences, convention d'échanges de données...)
- Préciser la (les) licence(s) pour la réutilisation de chaque jeu de données (Licences Creative Commons, Open Data Commons, License ouverte (Etalab), Tous droits réservés, autres...)
  - Bases de données : Licence ODBL 1.0
  - Licence ouverte pour la plupart des données : licence CC-BY-NC (selon partenaires et conventions)

Voir les 6 licences CC-BY

Organisation et documentation des données

Quelles 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

Origine et flux des données de l'UMR : Origine et flux de données UMR Biogeco
La cellule qualité de l'unité a mis en place un site sharepoint "GED : Gestion Electronique de Documents",
listant entre autres des modes opératoires et des protocoles, mis à disposition de tous les expérimentateurs terrains et laboratoires.
L'outil web de saisie des métadonnées pour les jeux de données est accessible à cette adresse :Outil de saisie des métadonnées pour les jeux de données.
Il est également prévu en 2020 de mettre à disposition une procédure permettant :

- de lister des règles communes de nommage pour faciliter et pérenniser l'accès à l'information
- de définir un plan de classement des dossiers

Etape	Méthode(s)	Outil(s)
Préparation	• Rédaction de mode opératoires et protocoles • Mise à disposition de tous les expérimentateurs terrains et plus largement de tous les agents de l'UMR.	• Traitement de texte • GED unité • OneDrive personnel (https://partage-fichiers.inra.fr/perso/LDAP)
Acquisition	• Modes opératoires et protocoles terrain sur le site sharepoint de l'unité • Pour les échantillons : collecte, conditionnement et identification, création fichier avec informations, alimentation de l'application **SAMPLES**, reconditionnement pour stockage.	• Saisie sur papier / ordinateur de terrain • Saisie des échantillons et leurs métadonnées dans bases de données SAMPLES • Outil de saisie des métadonnées pour les jeux de données
Traitement / Curation	• Convention de nommage pour les noms de fichiers et répertoires • Définition d'un plan de classement des répertoires • Métadonnées • Stockage sur support fiable des fichiers bruts (Poste de travail puis NAS unité) • Traçabilité des actions effectuées sur les données sur plusieurs supports: ◦ cahier de laboratoire (papier ou tableur) ◦ script de traitement (SAS, R, ...) ◦ NAS unité	• Excel, Libeoffice Calc, R, CSV, Système de Gestion de Bases de Données • Outil de saisie des métadonnées pour les jeux de données
Analyser les données	• Interprétation des données avec différents outils d'analyse (PGTB, GENOBOIS, Code informatique logiciels, ..) • Saisie des métadonnées concernant les outils utilisés : versions, paramètres, librairies, ...	• R, Programmation, QGis, cluster de calculs, etc... • Outil de saisie des métadonnées pour les jeux de données
Livraison et structuration jeux de données	• Stockage dans NAS unité • Intégration en base de données • Systèmes d'information	• Outil de saisie des métadonnées pour les jeux de données (ajout/modification)
Mise à disposition	Accès aux données via des applications Web dédiées ou clients compatibles, des entrepôts de données, des forges pour le code. Creation de scripts d'export pour mise à disposition vers des partenaires.	• **Interne UMR** ◦ Accès aux données via des applications Web dédiées ou des clients compatibles (R, Qgis, shiny...) ◦ Bases de données ◦ Systèmes d'information • **Externe UMR** ◦ Bases de données ◦ Systèmes d'information ◦ Entrepôt institutionnel DATA INRAE ◦ Entrepôt thématique ◦ Forge logicielle
Archivage	Pour les bases de données, les sauvegardes sont déplacées dans des espaces de stockage en interne, mais seront d'ici fin 2019 gérées par la DSI INRAE, dans un datacenter. Les jeux de données sont archivés le temps indiqué par le propriétaire	Outil d'interrogation des métadonnées pour les jeux de données Stockage réseau / Datacenter / Outils de transferts de données
Réutilisation	Nouvelles recherches, enseignement, réexamination des résultats, méta-analyses	Outil d'interrogation des métadonnées pour les jeux de données

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

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*
Ontologie des plantes ligneuses (mesures de traits phénotypiques)	saisie manuelle sur cahier de laboratoire ou ordinateur de terrain, annotation atomatique	Woody Plant Ontology	
Life Sciences Metadata	saisie manuelle sur cahier de laboratoire ou ordinateur de terrain, annotation atomatique	Vocabulaires contrôlés pour différentes métadonnées (sous-ensemble de OBI Ontology et NCBI Taxonomy for Organisms)	
Geospatial Metadata	saisie manuelle sur cahier de laboratoire ou ordinateur de terrain, annotation atomatique	?	

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

Mise en place de fichiers README pour chaque jeu de données + outil web de production de métadonnées. Au minimum, déposer dans un répertoire du NAS de l'unité le fichier de métadonnées généré par l'outil.

A minima, un fichier de type README pour lister les informations de base sur les données / bases de données / logiciels stockés : source, format du fichier, logiciel et traitements utilisés, identifiant, description du contenu, URL...

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

De manière générale, pour les fichiers, nous préconisons de suivre les règles de nommage indiquées sur le siteDORANUM
Nous nous reposons également sur le guide traçabilité des activités de recherche et gestion des connaissances pour rédiger une procédure sur le nommage et l'organisation des documents dans l'espace dédié aux jeux de données.

- Pour les échantillons, le choix se porte sur l'application web / Base de données**SAMPLES**, développée par l'INRA Orléans. Cet outil de gestion des échantillons est mis en place depuis Mars 2021.
- Pour les photos, une photothèque a été mise à disposition, accessible depuis un espace NAS 147.100.113.73 ou directement à l'adresse suivante:Albums photos

Quelle est la procédure de contrôle qualité des données ? joindre éventuellement le plan d'assurance qualité

Saisie de données
Le Gestionnaire Electronique de Documents (GED) donne accès à un certain nombre de procédures et de modes opératoires dans les domaines suivants :

- Laboratoires : Entomologie, Mycologie-Pathologie, Métrologie, Biologie moléculaire (Exemple : protocoles d'utilisation de matériels, extractions d'ADN, milieux de cultures, ...)
- Protocoles Terrain (récolte, conditionnement)
- Procédures informatiques
- Métrologie / Instrumentation

Systèmes d'informations / Bases de données
Concerne les SI et BDD suivants : Quercus portal, Pinus portal, QuercusMap DB, CMAP DB, TREEPOP DB, GD² DB, OAK PROVENANCE DB, SYLVCOOP DB, PINELINE DB

- Récupération des fichiers de soumission envoyés par les utilisateurs suite à la demande d'intégration (mise à disposition de fichiers modèles pour éviter l'hétérogénéité des formats et des variables)
- Copie des fichiers de soumission dans les répertoires de travail et les répertoires de sauvegarde de fichiers originaux
- Vérification générale du format soumis (nombre de lignes et colonnes respectées, aspect général des données saisies, présence de caractères spéciaux, cohérence des données saisies, respect des vocabulaires/ontologies)
- Insertion des données
- Vérification par les utilisateurs
- Validation finale en production
- Mise à disposition utilisateurs / pilier Forêt RARe / SI national GnpIS

Pilier Forêt RARe
Dans le cadre du pilier Forêt (Infrastructure de Recherche nationale : Ressources Agronomiques pour la Recherche - RARe), l'UMR Biogeco participe activement à l'enrichissement des collections de ce pilier avec de nouvelles ressources biologiques forestières (Introduction de RBF, Caractérisation des RBF, Gestion des SI). Le processus Systèmes d'information du pilier repose sur plusieurs objectifs, dans lesquels les SI de l'UMR Biogeco sont impliqués :

- améliorer la visibilité nationale et internationale des collections du pilier Forêt en accédant facilement aux métadonnées et données disponibles
- sécuriser les données
- intégrer des données de qualité

Les données des SI de l'UMR Biogeco (Quercus portal, Pinus portal) identifiées comme étant valides pour une remontée dans le catalogue de ressources RARe font l'objet à la fois des procédures Qualité de l'UMR et celles du processus SSI du pilier Forêt. Ces procédures concernent essentiellement l'intégration de données dans les bases de données locales et leur sauvegarde sur un espace sécurisé dédié au pilier Forêt (voir ici l'espace sharepoint du processus SSI du pilier).
Les données concernées par une remontée dans le pilier Forêt concernent des ressources génétiques et leurs données associées. Par exemple des arbres pour lesquels nous disposons :

- de matériel biologique (graines, pollens, plants)
- de données passeports ou métadonnées
- de données de génotypage & phénotypage
- de toute donnée de caractérisation du matériel disponible

La finalité est la mise à disposition des ressources biologiques et de leur données associées (rôle du SI global Forest tree GnpIS) afin de faciliter la valorisation et la distribution (via le panier de commande) des ressources biologiques forestières du Pilier Forêt.
Pour plus de détails, voir ici la fiche d'identité du processus SSI (Gestion des Systèmes d'information)

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)

- Stockage des données / Emplacement
  - Acquisition, nettoyage et curation des données : poste de travail (sauvegarde quotidienne des postes) puis NAS d'unité
  - Gestion / Partage des données au sein de l'UMR : NAS d'unité, Bases de données
    - Voir les Systèmes d'information et Base de données de l'UMR
  - Partage des données externe : Serveurs, Bases de données, Entrepôts de données (Data INRAE, NCBI, ...), Forges logicielles (SourceSup)
- Plan de sauvegarde des données
  - Bases de données : 2 copies à 2 endroits différents (sauvegardes automatisées) - Responsable : François Ehrenmann
    - pilier Forêt RARe : sauvegarde mensuelle sur serveur (procédure disponible sur le GED, section Informatique)
  - Transfert de fichiers entre ressources et espaces de stockage

Quelle est la volumétrie actuelle et prévisionnelle ?

Volumétrie actuelle des bases de données de l'UMR ~ 20 GB
Volumétrie des données brutes, données échantillons, scripts, programmes et documentations à estimer.

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...

Politique de sécurité en vigueur

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

Cela implique pour l'UMR :

- Sauvegarde quotidienne des postes de travail
- Sauvegarde quotidienne des bases de données (2 emplacements distincts)

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 ?

Droits à définir selon la nature des partenaires.
Avant publication, les jeux de données déposés sur l'espace réseau de l'UMR (NAS d'unité) sont sous la responsabilité du dépositaire des données, et les droits d'accès sont également décidés par ce dernier.
Les bases de données sont accessibles uniquement en lecture pour la totalité des utilisateurs, via des interfaces web. Les extractions de données lors de demandes

particulières sont réalisées par l'administrateur de bases de données.

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

SECURITE

- Protection contre les virus et les intrusions.
- Restrictions sur le droit d'accès (authentification nécessaire pour y accéder)
- Encryptage des données sensibles (mots de passe, données géographiques)
- Clause de confidentialité

INTEGRITE
Les procédures d'insertion en bases de données contrôlent un ensemble de critères et règles pour assurer la cohérence et l'unicité des données. Les jeux de données sont intégrés dans les bases par l'administrateur, après réception et contrôle des fichiers envoyés par les utilisateurs, grâce à des modèles.
TRACABILITE

- Bases de données : informations disponibles lors de l'insertion de données (date, propriétaire, type de jeu de données)
- Données brutes : responsable du jeu de données, dates de mise à jour des données et métadonnées
- Echantillons : informations disponibles lors de l'insertion des métadonnées des échantillons (date, propriétaire, projet, collecteur)

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 ?

Toutes les données générées par l'UMR ne sont pas destinées à être conservées à perpétuité (notamment les données traitées ou analysées). Seules les données issues d'acquisition et les données brutes, la plupart du temps non reproductibles, sont conservées sur le long terme.
Pour les données à durée de vie limitée, une politique de gestion des données concernant leur effacement permet d'utiliser plus efficacement l'espace de stockage disponible et permet de réduire le volume de métadonnées associées. Cette réduction permet aussi de réduire le temps nécessaire à localiser les données d'intérêt. Les réponses aux questions ci-dessous permettront de prendre une décision sur la durée de conservation des données.
Questions à poser aux détenteurs de données

- Envisagez-vous d'effectuer la préservation à long terme de vos données c'est à dire au-delà de l'arrêt de votre projet ? oui/non
- Que faire des données stockées pendant le projet mais qui ne seront pas archivées ? Faut-il envisager leur effacement/destruction ? Faut-il envisager un délai de grâce avant destruction définitive de celles-ci ?
- Quelle sera la durée de préservation des données au-delà du projet ?
- Pour les données archivées, quels sont vos projets pour une éventuelle transformation de leur format dans le futur ?

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 ?

- Plateforme spécifique d'Archivage pérenne au CINES ? (Convention à rédiger)
- Datacenter ? (espace de stockage institutionnel géré par la DSI de l'INRAE)
- Métadonnées sauvegardées sur site et sous Dataverse INRAE.

Quelle est la durée de conservation des données ?

A définir selon la nature des données et la nature des projets.

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

Fonds propres de l'unité et fonds provenant de projets de recherche.