
PGD UMR BIOGECO

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	PGD UMR BIOGECO
Version	4
Domaines de recherche (selon classification de l'OCDE)	Agriculture, forestry, and fisheries
Langue	fra
Date de création	2020-03-27
Date de dernière modification	2021-09-02
Identifiant	PGD-BIOGECO

Renseignements sur le projet

Titre du projet	PGD UMR BIOGECO
Résumé	<p>Plan de gestion de données de l'UMR Biogeco.</p> <p>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.</p>

Produits de recherche :

1. Ensemble des données générées et gérées par l'UMR Biogeco

Contributeurs

Nom	Affiliation	Rôles
Ehrenmann François	INRAE	<ul style="list-style-type: none"> • Personne contact pour les données • Responsable du plan de gestion de données
Chaumeil Philippe	INRAE	
Plomion Christophe	INRAE	<ul style="list-style-type: none"> • Coordinateur du projet

Droits d'auteur :

Le(s) créateur(s) de ce plan accepte(nt) que tout ou partie de texte de ce plan soit réutilisé et personnalisé si nécessaire pour un autre plan. Vous n'avez pas besoin de citer le(s) créateur(s) en tant que source. L'utilisation de toute partie de texte de ce plan n'implique pas que le(s) créateur(s) soutien(nen)t ou aient une quelconque relation avec votre projet ou votre soumission.

Informations sur la structure

UMR1202 BIOGECO Biodiversité, Gènes et Communautés

- 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>

Dataverse : <https://data.inra.fr/dataverse/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 Hérité.

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.

Origine 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)
- code informatique (logiciels, bases de données)

Nature 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

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
Emmanuel Corcket	emmanuel.corcket@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

-
- INRAE
 - Université de Bordeaux

-
- EFPA

Nouvellement ECODIV

- [Site web](#)
- [Intranet](#)

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

Informations sur le plan de gestion

Date	n° de version	Status	Auteur	Affiliation de l'auteur (se reporter à l' annuaire Inra)	Validé par	Validé le
15/12/2020	3	Validé	<ul style="list-style-type: none"> François Ehrenmann Philippe Chaumeil 	UMR 1202 BIOGECO	Christophe Plomion	15/12/2020
06/09/2019	1	Non validé	<ul style="list-style-type: none"> François Ehrenmann Philippe Chaumeil 	UMR 1202 BIOGECO		
26/09/2019	1	Non validé	<ul style="list-style-type: none"> François Ehrenmann Philippe Chaumeil 	UMR 1202 BIOGECO		
19/11/2019	1	Non validé	<ul style="list-style-type: none"> François Ehrenmann Philippe Chaumeil Christophe Plomion 	UMR 1202 BIOGECO		
12/03/2019	2	Non validé	<ul style="list-style-type: none"> François Ehrenmann Philippe Chaumeil Christophe Plomion 	UMR 1202 BIOGECO		

Présentation générale des données

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

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

Science participative

-
- Software
 - Model
 - Image
 - Dataset
 - Collection
 - 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

-
- 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)

Spécifiez 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](#)

Plusieurs 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

-
- Soils and soil sciences
 - Plant Health and Pathology
 - Plant Breeding and Plant Products
 - Omics
 - Insects and Entomology
 - Forests and Forest Products

- Biodiversity and Ecology
-

Droits de propriété intellectuelle

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 issu des travaux de l'unité doit être publié en libre accès : (i) via la voie diamant (gratuité totale), (ii) la voie verte (1. auto-archiver d'un preprint non évalué par les pairs, par exemple dans BIORXIV, 2. archiver 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

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](#)

- 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.
-

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

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.

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

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

- **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)
 - Espaces 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. L'UMR a son propre dataverse sur le portail Data INRAE : [Dataverse Biogeco](#)
 - 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 public 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](#)

- 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).

- 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, Licence 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

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	<ul style="list-style-type: none">• Rédaction de mode opératoires et protocoles• Mise à disposition de tous les expérimentateurs terrains et plus largement de tous les agents de l'UMR.	<ul style="list-style-type: none">• Traitement de texte• GED unité• OneDrive personnel (https://partage-fichiers.inra.fr/perso/LDAP)
Acquisition	<ul style="list-style-type: none">• 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.	<ul style="list-style-type: none">• 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	<ul style="list-style-type: none">• 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:<ul style="list-style-type: none">• cahier de laboratoire (papier ou tableur)• script de traitement (SAS, R, ...)• NAS unité	<ul style="list-style-type: none">• 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	<ul style="list-style-type: none">• Interprétation des données avec différents outils d'analyse (PGTB, GENOBOIS, Code informatique logiciels, ..)• Saisie des métadonnées concernant les outils utilisés : versions, paramètres, librairies, ...	<ul style="list-style-type: none">• R, Programmation, QGIS, cluster de calculs, etc...• Outil de saisie des métadonnées pour les jeux de données
Livraison et structuration jeux de données	<ul style="list-style-type: none">• Stockage dans NAS unité• Intégration en base de données• Systèmes d'information	<ul style="list-style-type: none">• 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](#)

Métadonnées	Origine, mode de production des métadonnées (ex : saisie manuelle, annotation automatique...)	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 automatique	Woody Plant Ontology	
Life Sciences Metadata	saisie manuelle sur cahier de laboratoire ou ordinateur de terrain, annotation automatique	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 automatique	ontologie des géolocalisation ?	

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...

Réflexion se reposant sur guide [traçabilité des activités de recherche et gestion des connaissances](#) pour

rédigier une procédure sur le nommage et l'organisation des documents dans l'espace dédié aux jeux de données.

- Pour les échantillons, le choix se portera sur l'application web / Base de données **SAMPLES**, développée par INRA Orléans. Cet outil de gestion des échantillons sera mis à disposition au premier trimestre 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](#)

Gestion de version

- Fichiers de données : à chaque modification, la partie version du nom du fichier est incrémentée.
- Code informatique (applications web, logiciels, scripts) : GIT / GIT-LAB

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

- 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 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).

Stockage et sécurité des données

- 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

- Transfert de fichiers entre ressources et espaces de stockage

Volumétrie actuelle des bases de données de l'UMR : 3 Go

Volumétrie des données brutes, données échantillons, scripts, programmes et documentations à estimer.

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)
-

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

- 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

INTÉGRITÉ

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.

TRAÇABILITÉ

- 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
 - Échantillons : informations disponibles lors de l'insertion des métadonnées des échantillons (date, propriétaire, projet, collecteur)
-

Archivage et conservation des données

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 ?

-
- 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](#).

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

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