Plan de Gestion de Données (modèle INRAE structure) de la Plateforme d'Exploration du Métabolisme – MetaboHUB Clermont Ferrand (PFEM)

Plan de gestion de données créé à l'aide de DMP OPIDoR, basé sur le modèle "INRAE - Modèle Structure (v2)" fourni par INRAE - Institut national de recherche pour l'agriculture l'alimentation et l'environnement.

Renseignements sur le plan

Titre du plan	Plan de Gestion de Données (modèle INRAE structure) de la Plateforme d'Exploration du Métabolisme – MetaboHUB Clermont Ferrand (PFEM)
Livrable	
Version	Version initiale
Objet/périmètre du plan	Ce document présente le **Plan de Gestion de Données de structure de la Plateforme d'Exploration du Métabolisme** – MetaboHUB Clermont Ferrand (PFEM). Il a également pour objet de servir de support à l'écriture des plans de gestion de données des projets dans lesquels la PFEM est partenaire.
Domaines de recherche (selon classification de l'OCDE)	1.4 Sciences chimiques, 1.6 Sciences biologiques, 1.2 Informatique et sciences de l'information
Langue	fra
Date de création	2024-06-06
Date de dernière modification	2024-06-12
Identifiant	10.5281/zenodo.11612926
Type d'identifiant	DOI
Licence	**Nom** Creative Commons Attribution 4.0 International **URL** http://spdx.org/licenses/CC-BY-4.0.json
Documents (publications, rapports, brevets, plan expérimental....), sites web associés	• Portail web de la PFEM : https://pfem.isc.inrae.fr/ • Identifiant Institutionel de la PFEM : 10.15454/1.557239511356972E12
Plans de gestion en lien avec le projet	• PGD du projet AgroDataRing :

Détails de la Structure

Nom de l'entité	Plateforme d'Exploration du Métabolisme
Acronyme	PFEM
Identifiant	10.15454/1.557239511356972E12
Type d'identifiant	URL
Description	La PlateForme d’Exploration du Métabolisme (PFEM), est une infrastructure scientifique et technique qui a pour objectif de mettre à disposition de la communauté scientifique publique et privée les concepts, méthodes et outils pour l’étude du métabolisme. Partenaire dans les projets de recherche qu’elle accompagne, elle possède une expertise dans les domaines de la nutrition, santé, environnement et qualité des produits alimentaires. La PFEM est structurée en deux composantes : • Métabolomique, adossée à l’Unité de Nutrition Humaine (UMR 1019 INRAE UCA) et à l’Institut de Chimie de Clermont-Ferrand (UMR6296 CNRS UCA) • Protéomique, adossée à l’unité INRAE Qualité des Produits Animaux (UR0370) La PFEM est labellisée par le GIS IBiSA depuis 2010 et certifiée selon les normes ISO 9001 et NFX 50-900, et sa composante métabolomique est intégrée dans l’infrastructure nationale de métabolomique et fluxomique MetaboHUB, avec les plateformes de Bordeaux, Toulouse, Paris-Saclay, Rennes-Nantes.
Date de création	2010-01-01
Tutelle(s) de l'entité	• National Research Institute for Agriculture, Food and Environment https://ror.org/003vg9w96 • University of Clermont Auvergne https://ror.org/01a8ajp46
Source(s) de financement	• Agence Nationale de la Recherche : • European Commission :

Produits de recherche :

1. Métabolomique non ciblée (Service)
2. Métabolomique semi ciblée (Service)
3. Fluxomique ciblée (Service)
4. Identification structurale (Service)
5. Analyses statistiques (Workflow)
6. Bioinformatique (Logiciel)
7. Quantification Label Free (Service)
8. Quantification ciblée (Service)
9. Identification (Service)
10. Profils MALDI-TOF (Service)

Contributeurs

Nom	Affiliation	Rôles
Chambon Christophe		• Personne contact pour les données (Proteo-PR07, Proteo-PR09, Proteo-PR08, Proteo-PR10) • Responsable de l'entité
Durand Stéphanie		• Responsable de l'entité
Giacomoni Franck - 0000-0001-6063-4214	INRAE - 039gscz82	• Personne contact pour les données (Metabo-PR06) • Responsable du plan • Responsable du plan
Hebraud Michel		• Responsable de l'entité
Paulhe Nils		
Pétéra Mélanie		• Gestionnaire de données • Personne contact pour les données (Metabo-PR05)
Pujos-Guillot Estelle		• Personne contact pour les données (Metabo-PR03, Metabo-PR04, Metabo-PR01, Metabo-PR02) • Responsable de l'entité

Plan de Gestion de Données (modèle INRAE structure) de la Plateforme d'Exploration du Métabolisme – MetaboHUB Clermont Ferrand (PFEM)

Informations sur la structure

Nom de la structure

Plateforme d’Exploration du Métabolisme (PFEM)

Type de structure

- Plateforme, plateau technique
- Plateforme, plateau technique

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

https://hal.inrae.fr/PFEM

Responsabilités dans la structure

Nom, Prénom	Courriel	Rôle
Pujos-Guillot Estelle		Responsable scientifique
Hebraud Michel		Responsable scientifique
Durand Stéphanie		Responsable technique
Chambon Christophe		Responsable technique

Etablissement(s) tutelle(s)

INRAE

Département de rattachement INRAE (ou anciens départements Inra)

- ALIM-H : Alimentation humaine
- TRANSFORM : Aliments, produits biosourcés et déchets
- ALIM-H : Alimentation humaine
- TRANSFORM : Aliments, produits biosourcés et déchets

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

INRAE

Informations sur le plan de gestion

DOI (version publiée du plan de gestion)

https://zenodo.org/doi/10.5281/zenodo.11612926

Historique des versions

Date	n° de version	Status	Auteur	Affiliation de l'auteur *(se reporter à l'annuaire INRAE)*	Validé par	Validé le
12/2022	Version 0.1		C. Dupérier, M. Pétéra et F. Giacomoni	INRAE UNH PFEM	E Pujos-Guillot	02/2023
02/2023	Versions 0.2 & 0.3		C. Dupérier, M. Pétéra et F. Giacomoni	INRAE UNH PFEM	E Pujos-Guillot	03/2023
04/2023	Versions 0.4 et 0.5		Groupe de travail PFEM	INRAE UNH PFEM	E Pujos-Guillot	07/2023
07/2023	Version 0.6		Groupe de travail PFEM	INRAE UNH PFEM	E Pujos-Guillot	05/2024
06/2024	Version 0.7		C. Dupérier, M. Pétéra et F. Giacomoni	INRAE UNH PFEM	E Pujos-Guillot	06/2024
06/2024	Version 1.0		C. Dupérier, M. Pétéra et F. Giacomoni	INRAE UNH PFEM	E Pujos-Guillot	06/2024

Présentation générale des données

Mode d'obtention des données

- Données générées par la structure
- Données générées par la structure
- Projets incluant l'analyse de matériel biologique

Génération de la donnée brute :

Les machines d'analyses de type spectromètres de masse (MS) et de type spectromètres de résonance magnétique nucléaire (RMN) sont pilotées par des ordinateurs d'acquisition via des logiciels constructeurs qui vont produire les données brutes analytiques :

Pour la MS, ce sont des chromatogrammes, dont les spectres seront issus. Ces données seront disponibles sous deux types de formats de fichiers pouvant être produits en parallèle. :

Fichier brut, issu des formats constructeurs, souvent encodés. Exemple : fichier avec extension *.d;
Fichier au format ouvert, converti depuis le fichier brut. Exemple : fichier avec extension *.NetCDF ou *.mzdata.

Pour la RMN, ce sont par exemple des « FID ». Ces données seront par exemple disponibles sous le format « 1r ».

Caractéristiques de ces données brutes :

Dépendantes du logiciel pour l'analyse (constructeur)
Contiennent une liste de séquences (échantillons, blancs, pools, …)
Une méthode unique pour tous les échantillons décrivant les conditions d'analyses
Calibration et réglages de l'appareil
Structurée et organisée en répertoires et fichiers conditionnant l'intégrité de la donnée brute selon les constructeurs

Points de vigilances :

Le format des données brutes évoluant avec les versions des logiciels constructeurs, la PFEM porte une attention particulière au suivi de ces évolutions.

- Projets de codes informatiques et bases de connaissances

Dans ce document, les logiciels, programmes et scripts produits par la PFEM seront nommés "codes informatiques", et les bases-de-données et les réseaux de connaissances sont désignés comme « bases de connaissances ».

Génération de codes informatiques :

La génération des codes est effectuée sous différents langages de programmation choisis en fonction du besoin (R, Java, Python, Perl, PHP, …), directement sur les machines des ingénieurs membres du projet, sous des environnements de développement intégré (Exemple : Eclipse, RStudio, IntelliJ, …).

Point de vigilances :

Documentation et suivi de version sur la forge PFEM
Sécurisation des codes exposés sur le Web
Maintenance des codes (mise à jour, résolution de bogues, …)

Origine

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

Type de données

- Dataset
- Software
- Workflow
- Dataset
- Software
- Workflow

Le plan distingue plusieurs typologies de données :

· Des données collectées : publications, données de l'état de l'art du domaine (exemple avec les bases de données) ou données externes de collaborateurs (métadonnées d'échantillons).

· Des données produites par la PFEM, soit à l'issue d'une analyse chimique, soit à l'issue d'un traitement de données d'analyse.

· Des codes informatiques considérés également comme données produites dès qu'ils sont publiés sur des dépôts publics.

· Des donnée traitées : un ensemble des fichiers et répertoires obtenus à partir des données d'acquisition et modifiés à

l’aide d’un algorithme ou d’une chaîne de traitement.

Nature des données

Question sans réponse.

Format des données

· Fichier brut, issu des formats constructeurs, souvent encodés. Exemple : fichier avec extension **.d*;
· Fichier au format ouvert, converti depuis le fichier brut. Exemple : fichier avec extension *.NetCDF ou *.mzdata.

DataTypes (RAW)

code	name	version	comments
_D_BRUKER_V4.3	format .D (Bruker) v4.3.100	v4.3.100	from Bruker Impact on Data Analysis
_D_BRUKER_V3.4	format .D (Bruker) v3.4	v3.4	from Bruker AutoFlex on Flex Analysis
_D_AGILENT_V7.05	format .D (Agilent) vB.07.05.2479	vB.07.05.2479	from GCTOF on MassHunter
_D_AGILENT_V7.00	format .D (Agilent) VB.07.00.7024	vB.07.00.7024	from Agilent13 on MassHunter
_D_AGILENT_V7.01	format .D (Agilent) VB.07.01.1805	vB.07.01.1805	from Agilent14 on MassHunter
_RAW_THERMO_V2.2	format .raw (Thermo) V2.2	v2.2 SP1.48	from Orbitrap and LTQ on XCalibur
_RAW_THERMO_V2.07	format .raw (Thermo) V2.07	v2.07 SP1	from Quantum on XCalibur
_RAW_WATERS_V4.1.627	format .raw (Waters) V4.1 SCN627	v4.1 SCN627	from LC-QqQ-Quattro on Masslynx 4.1 SCN627
_RAW_WATERS_V4.1.805	format .raw (Waters) V4.1 SCN805	v4.1 SCN805	from GC-QqQ-Quattro on MassLynx 4.1 SCN805
_RAW_GVINSTR_V1.4	format .raw (GV Instr.) V1.4	v1.4	from gc-c-irms on IonVantage 1.4
_WIFF_ABSCIEX_V1.6	format .wiff (ABSciex) v1.6.2	v1.6.2	from QTRAP on Analyst
_WIFF-SCAN_ABSCIEX_V1.6	format .wiff.scan (ABSciex) v1.6.2	v1.6.2	from QTRAP on Analyst

DataTypes (Converted)

code	name	version	comments
_CDF_V2.3.2	NETCDF v2.3.2	v2.3.2	Bruker Data analysis 4.3.100 autogenerated NETCDF
_MZDATA_V1.05	MZDATA v1.05	v1.05	Agilent MassHunter for QTOF 7.01 and QQQ autogenerated MZData
_MZXML_V3.2	MZXML	v3.2	ABSciex Analyst for Qtrap 1.6.2 generated mzXML with ProteoWizard software version=3.0.9248

Périmètre thématique des données

- Computer science
- Food and food processing
- Human Health and Pathology

- Human Nutrition and food security
- Information management
- Omics
- Computer science
- Food and food processing
- Human Health and Pathology
- Human Nutrition and food security
- Information management
- Omics

Droits de propriété intellectuelle

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

En tant que plate-forme membre de l'infrastructure nationale MetaboHUB, les aspects juridiques (titularité ou droits de propriété intellectuelle sur les données) sont traités et validés à travers l'accord de consortium de cette infrastructure.

Les aspects juridiques (titularité ou les droits de propriété intellectuelle sur les données enrichies et/ou valorisées) sont traités et validés à travers l'accord de consortium de l'infrastructure MetaboHUB (Avenant n°5 – 2022). La propriété intellectuelle des données enrichies et/ou valorisées est discutée avec le porteur du projet au montage du projet scientifique demandeur des analyses et formalisés dans les contrats signés avec les Tiers. Cette ouverture des données sera définie en concertation avec le porteur du projet scientifique demandeur des analyses mais une période d'embargo devra être clairement définie dans le contrat ou l'accord de consortium du projet avec le Tiers pour permettre leur éventuelle publication sur des dépôts de référence en lien avec les publications des résultats scientifiques obtenus par l'analyses et l'interprétation de ces données. Il est à noter que selon la loi n°2016-1321 pour une République numérique du 7 octobre 2016 prévoit qu'une donnée sera qualifiée de libre si cette donnée est issue d'une activité de recherche financée au moins pour moitié par des fonds publics et que ces données ne sont pas protégées par un droit spécifique et que ces données ont été rendues publiques par le chercheur ou l'établissement.

Un document interne du système documentaire de la PFEM liste les différents types de projets et les cas de propriétés intellectuelles rattachées.

Sensibilité des données

Identification du niveau de sensibilité des jeux de données

- Public
- Diffusion limitée
- Confidentiel
- Public
- Diffusion limitée
- Confidentiel

La PFEM se positionne en faveur de la science ouverte. Les principales contraintes au partage de données résident dans le respect des dispositions légales, en particulier en cas de confidentialité des données liée aux protocoles de recherche des partenaires (données sensibles, personnelles ou critiques). Dans ce cas, les conditions spécifiques au projet de recherche sont définies en amont du projet avec le partenaire (garant du respect de la législation), et les contraintes sont explicitement décrites. Communément, ces informations sont décrites dans le contrat de recherche du projet du partenaire.

Dans le cas de production d'algorithmes et codes sources, en cas de suspicion d'innovation (au sens tel que défini par INRAE et UCA), le partage des données peut être mis en attente le temps de l'évaluation du caractère innovant par les commissions compétentes.

Les données de la PFEM à la vocation à devenir publiques sont publiées sous licence de type Creative Commons. En dehors de cas spécifiques liés à des licences contaminantes ou des contrats de recherche, les codes informatiques produits seront eux généralement publiés sous deux types de licences : MIT ou CeCILL 2.0.

Quelles sont les mesures prises et les normes auxquelles il est nécessaire de se conformer pour garantir la sécurité des données sensibles ?

Le stockage des données est un service proposé aux usagers de la PFEM dans le cadre des collaborations scientifiques. Les conditions de ce stockage se font en accord avec les contraintes détaillées dans le contrat de recherche passé entre la PFEM et le porteur du projet. La PFEM s'engage notamment à assurer une sécurisation des données (accès, sauvegarde, ...) à l'état de l'art ainsi qu'à leur conservation sur des pas de temps alignés avec la législation (Exemple : 10 ans de conservation des données humaines). A noter que la PFEM n'assure pas et ne propose pas de service d'archivage long terme de données (supérieur à 10 ans).

S'il y a des données à caractère personnel, quelles sont les mesures envisagées pour les protéger au cours du projet ou dans le cadre d'une réutilisation ?

Les données administratives des projets sont recueillies par l'outil de gestion de projets dédié ou transmises par le porteur du projet. Ce recueil a comme objectifs i) d'évaluer les demandes et permettre les contacts avec les demandeurs, ii) d'assurer la traçabilité du projet, iii) de calculer des indicateurs internes à la PFEM en lien avec la politique qualité de la plate-forme et iv) de calculer des indicateurs globaux remontés aux les tutelles et aux financeurs de la PFEM.

Les objectifs du recueil de données sont explicités via le portail web d'accès aux services, en conformité avec la réglementation européenne notamment en lien avec le Règlement Général sur la Protection des Données (RGPD). Dans le cas des données administratives purement internes à la PFEM, celles-ci sont gérées par le système de gestion de la qualité de la plate-forme selon les procédures définies et accessibles, en interne, à tous les personnels PFEM. Une action est en cours pour déployer les recommandations définies par INRAE de mise en conformité RGPD de ses plateformes.

Le stockage des données est un service proposé aux usagers de la PFEM dans le cadre des collaborations scientifiques. Les conditions de ce stockage se font en accord avec les contraintes détaillées dans le contrat de recherche passé entre la PFEM et le porteur du projet.

Partage des données

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

Le partage public de données est un élément clé dans l'accélération de la recherche et un des piliers d'une science plus ouverte. Il permet notamment la réutilisation de données pour une nouvelle question scientifique, la comparaison d'études, l'annotation dans le cadre de d'autres projets, l'entrainement d'algorithmes, ou la conduite de meta-analyses. Sauf cas très particulier en lien avec un accord de consortium ou un contrat de recherche, La PFEM en tant que structure académique, souhaite s'inscrire dans la dynamique impulsée par le MESRI, les agences de financement, les établissements de recherche et les universités en termes de données ouvertes. Selon les textes en vigueur actuellement, deux conditions préliminaires sont actuellement nécessaires pour diffuser les données selon ces principes : des données réalisées dans le cadre de la mission de service public des établissements et des données achevées.

Ce partage public de données concerne alors les données brutes de standards chimiques, les données d'échantillons transformées en format ouvert. Cela concerne également les données enrichies et les workflows de traitement. Les données seront partagées avec la publication des résultats de l'étude par le porteur du projet, tel que défini dans le contrat avec ce tiers. Si la période d'embargo définie dans l'accord de consortium est dépassée, une réévaluation de l'ouverture sera effectuée au cas par cas en concertation avec le porteur.

La PFEM peut actuellement fournir un accompagnement en termes de moyens au porteur du projet scientifique. Les moyens mis à disposition seront discutés au montage du projet. Techniquement, le PFEM propose un espace publique accessible depuis internet (*via* une adresse web unique) vers les données brutes converties en format ouvert.

Quelles sont les réutilisations potentielles de ces données ?

La PFEM a pour objectif de mettre à disposition de la communauté scientifique publique et privée les concepts, méthodes et outils pour l'étude du métabolisme. L'objectif de la collecte et de la production de données est donc soit la mise au point de nouvelles méthodes d'analyses ou de nouveaux outils de traitement ou d'interprétation de données (activité R&D), soit la production de connaissance dans les différents domaines d'applications (activité routine, expertise etc.). De plus, la PFEM peut être amené à réutiliser des données de recherche notamment en vue de développer de nouvelles méthodes (R&D) et de produire de nouvelles connaissances en biologie.

La réutilisation de données au sein de la PFEM se fait en adéquation avec les licences des données ré-utilisées, conformément à la législation. Lorsque les données sont fournies par des partenaires, ceux-ci s'engagent à ce que la licence des données transmises soit compatible avec l'exploitation prévue des données.

Lorsque le partenaire produit lui-même ces données et qu'une licence particulière n'est pas déjà attachée à ces données, la PFEM recommande au partenaire l'utilisation d'une licence de type Creative Commons.

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

Les données brutes produites sont des formats de constructeurs, fermés et encodés. Des logiciels spécifiques à chaque constructeurs sont alors requis pour la lecture. Des solutions logiciels sont alors utilisées pour convertir ces formats en des formats ouverts et standardisés.

Comment les données seront-elles partagées ?

Ce partage public de données concerne alors les données brutes de standards chimiques, les données d'échantillons transformées en format ouvert. Cela concerne également les données enrichies et les workflows de traitement. La PFEM incite dès l'élaboration d'un projet scientifique à ouvrir les données du projet avec i) le dépôt des données instrumentales sur AgroDataRing et leur déclaration sur data.gouv.fr (Obtention de DOI) puis ii) le dépôt des métadonnées de l'étude sur MetaboLights ou Metabolomics Workbench (dépôts scientifiques de références). Les données seront partagées avec la publication des résultats de l'étude par le porteur du projet, tel que défini dans le contrat avec ce tiers. Si la période d'embargo définie dans l'accord de consortium est dépassée, une réévaluation de l'ouverture sera effectuée au cas par cas en concertation avec le porteur. Les données de la PFEM à la vocation à devenir publiques seront publiées sous licence de type Creative Commons sauf en cas de spécifications différente édictées par le contrat de recherche. Ce type de partage s'adresse à la communauté scientifique avec une demande préliminaire à adresser à la PFEM permettant de décrire le cadre de la réutilisation.

Avec qui ?

Question sans réponse.

Sous quelle licence ?

- Licence ouverte https://www.etalab.gouv.fr/licence-ouverte-open-licence (compatible CC-BY)
- Licence ouverte https://www.etalab.gouv.fr/licence-ouverte-open-licence (compatible CC-BY)

Les données de la PFEM à la vocation à devenir publiques seront publiées sous licence de type Creative Commons sauf en cas de spécifications différente édictées par le contrat de recherche. Ce type de partage s'adresse à la communauté scientifique avec une demande préliminaire à adresser à la PFEM permettant de décrire le cadre de la réutilisation.

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

- **Enregistrement et statut des projets via des outils de gestion de projets**

Chaque demande de projet faite à la plateforme est déposée par les utilisateurs sur les portails web de gestion de projets spécifiques aux projets de métabolomique et aux projets de protéomique.

Ces portails permettent de consulter l'ensemble des projets et de suivre ses différentes phases de l'état d'avancement. Chaque projet dispose d'un identifiant unique, utilisée dans l'ensemble des logiciels de traçabilité et de suivi de la PFEM.

- **Suivi des projets de Recherche & Développement PFEM**

Dès lors qu'un projet est accepté, un outil de gestion et de suivi macroscopique des projets assure la traçabilité et stockage des documents de suivi. Cet outil permet également le pilotage interne des projets et récupère automatiquement les informations renseignées sur les portails web d'enregistrement des demandes.

Lorsqu'un projet nécessite une analyse d'échantillons biologiques, alors il est enregistré automatiquement dans un système de gestion de l'information du laboratoire (LIMS). Ce LIMS permet le suivi des échantillons et des analyses. Un numéro unique est attribué à chaque échantillon à analyser. La nomenclature utilisée pour l'attribution des identifiants des échantillons contient les informations relatives à chaque client ainsi que les identifiants générés par les outils de gestion de projets.

- **Suivi des échantillons et de leur conformité**

A leur arrivée sur la plateforme, la conformité des échantillons est vérifiée et les informations sont consignées dans un cahier de réception physique. Ils sont ensuite stockés au -80°C jusqu'à l'analyse et les boites de stockage avant et après analyse sont renseignées dans le LIMS pour la LC-MS. Mais pour la GC-MS, les échantillons ne sont pas conservés après analyse à cause de phénomènes d'évaporation.

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 ?

La PFEM travaille activement dans le cadre de l'infrastructure MetaboHUB et du CATI EMPREINTE sur l'intégration de vocabulaires contrôlés dans les composants de son système d'informations au travers du projet MetaSaurus.

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

Les métadonnées des projets incluant l'analyse de matériel biologique, décrivant le contexte de l'étude (sources des échantillons, objectifs scientifiques, …) ainsi que les données qualitatives / quantitatives nécessaires à sa compréhension et à l'analyse des résultats seront mises à disposition par le porteur du projet aux responsables de la PFEM.

Les métadonnées des projets de codes informatiques et de bases de connaissances, décrivant le contexte (ressources, objectifs scientifiques, spécifications techniques,) ainsi que les données nécessaires à la constitution des bases seront mises à disposition par le(s) porteur(s) du projet aux chefs de projets PFEM et seront conservées sur la forge logicielle privée de la PFEM.

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

- **Nomenclature et arborescence utilisées**

Le système d'information de la PFEM permet de garantir la traçabilité des données au cours de leur cycle de vie,

notamment au moyen des identifiants LIMS associés à chaque échantillon et une conservation stricte des métadonnées des fichiers bruts (noms, dates, ...) depuis l'ordinateur d'acquisition, le serveur de stockage de la PFEM et jusqu'aux serveurs de traitement. Lors de l'analyse des échantillons, un nom est ainsi attribué à la donnée brute produite. Pour la MS, il contient des informations sur son identifiant LIMS, l'identifiant du client. Pour la RMN, les données sont nommées avec la nomenclature fournie par le demandeur.
Concernant l'organisation des données brutes produites en métabolomique, chaque stockage d'ordinateur de pilotage d'un instrument propose une arborescence structurée de répertoires et de fichiers sous le répertoire racine « Projets ». Pour chaque projet accepté, un sous-dossier est alors créé et son nom est constitué du numéro et nom du projet. Les données brutes y seront ainsi enregistrées et centralisées sur le serveur de stockage de la PFEM en conservant cette organisation. Cet espace contiendra également toute la documentation du projet (fichiers étiquettes), les informations de randomisation (ordre de passage des échantillons) voir les scripts du robot utilisés ordonné en sous dossiers. Toutes les données issues du pré-traitement (extraction des données) du traitement (filtres et normalisation), des statistiques, des annotations et des rendus de résultats sont aussi enregistrées sur le serveur de stockage de la PFEM sous le même dossier.

- **Cas des codes informatiques et bases de connaissances**

Les codes informatiques sont suivis grâce à l'utilisation de forges Logicielles. Ce type d'outils permet le stockage et la sécurisation du code (le code écrit sur une machine individuelle est centralisé sur un serveur PFEM), de gérer les accès des personnels et la publication des codes (visibilité privée ou interne ou publique), le suivi des versions du code et l'application de numéro de « releases » par un gestionnaire de suivi de code, de partager un même code entre membres du projet, de centraliser la documentation pour le développement et l'utilisation de ces codes informatiques, la vérification en continue de la qualité des codes (tests unitaires & fonctionnels) et d'automatiser leur(s) intégration(s) et leur(s) déploiement(s). La PFEM utilise plusieurs forges logicielles pertinentes en fonction des projets, notamment i) une instance GitLab pour les projets internes et privés à la PFEM, ii) le dépôt GitHUB "Workflow4Metabolomics" pour le déploiement des outils de traitement de données MS et RMN sur les plateformes permettant le traitement de données, iii) le dépôt GitHUB "eMetaboHUB" pour les outils publiés dans des journaux scientifiques et iv) le dépôt "MetaboHUB" sur le GitLab "INRAE" pour les projets collaboratifs entre équipes MetaboHUB.

Quel est le processus de contrôle qualité des données ?

L'ensemble du parc informatique (serveurs et machines individuelles) est sous maintenance et est contrôlé régulièrement par le responsable du système d'information de la PFEM. Dans ce contexte, la PFEM a mis en place plusieurs processus de vérification de la qualité de ses données et cela de leur génération à leur traitement.

- **Processus de vérification - Données analytiques**

En lien avec sa certification ISO-9001 et NFX-50900, la PFEM a mis en place différentes procédures et modes opératoires internes en LC-MS, GC-MS et RMN permettant d'assurer un contrôle efficace de la qualité des données analytique acquises. Il s'agit par exemple de la réalisation de calibration avant l'analyse des échantillons d'un projet, de l'utilisation d'échantillons de contrôle de la qualité, injectés régulièrement lors d'un projet pour vérifier la fiabilité de l'appareil.
Pour la spectrométrie de masse, les données brutes sont converties en fichiers au format ouverts et traitées sur une plateforme Galaxy. Concernant la RMN, les fichiers brutes sont traités grâce au logiciel NMRProcFlow. La visualisation des fichiers de sorties durant le traitement permet de contrôler la qualité des données. Un lien est généré pour chaque historique de traitement sous Galaxy, permettant ainsi la traçabilité et reproductibilité des analyses de données. Ce lien est reporté dans un document sauvegardé dans le dossier des données traitées.

- **Processus de vérification – Codes informatiques**

Le niveau demandé pour la qualité des codes produit au sein de la PFEM est fixé au travers de différents modes opératoires définissant les bonnes pratiques PFEM de développement pour différents langages de programmation utilisés. Cette qualité des codes informatiques est contrôlée par l'intégration obligatoire de tests unitaires et fonctionnels à chaque code. Ces tests sont ensuite régulièrement joués par l'intermédiaire de chaines d'intégration et de déploiement continues (CI/CD), configurées sur nos forges Logicielles. L'utilisation d'outil de gestion de tickets permet également un suivi fin et efficace de la prise en charge et de la gestion de problèmes détectés par les utilisateurs de ces codes informatiques.

Stockage et sécurité des données

Les systèmes d'information de la structure ont-ils fait l'objet d'une analyse de risques ou d'une homologation ?

- Oui
- Oui

La PFEM a conduit un audit d'homologation de son système d'information en 2022 en lien avec le responsable de la sécurité des systèmes d'information INRAE.

Quels types de supports physiques sont utilisés pour stocker les données ?

Les données expérimentales produites par la PFEM sont qualifiées de durables et non reproductibles. La plateforme sécurise toutes les données brutes produites en les stockant sur des serveurs informatiques adaptés, hébergés dans les salles de serveurs informatiques de l'institut. L'infrastructure maîtrise l'ensemble du cycle de vie de ses données par la mise en place d'outils modernes de gestion des données (stockage - sauvegarde - non archivage, génération de métadonnées pour les données brutes et enrichies, diffusion vers les partenaires et les référentiels) en accord avec les contraintes des accords de consortium des projets scientifiques auxquels la PFEM est associé.

En entrée de processus, les données brutes sont générées sur un poste d'acquisition pilotant l'instrument utilisé. Pour les instruments connectés au réseau INRAE, les données brutes sont alors copiées automatiquement au moins une fois par jour sur un serveur de stockage sécurisé, dans un espace dédié. Ce type de copie permet la synchronisation des fichiers de données brutes générés entre le poste d'acquisition et l'espace dédié sur le serveur de stockage, en conservant le nom et le chemin des fichiers. Pour les instruments non connectés au réseau et dont les machines d'acquisition disposent de systèmes d'exploitation trop anciens, la sauvegarde est manuelle et régit par une instruction du référentiel PFEM. Les données hébergées par la PFEM et les informations relatives à chaque le projet sont stockées sur un serveur de stockage sécurisé (double parité / disque de secours) permettant la reconstruction automatique des données en cas de défaillance matériel (perte de deux disques de données). Un service d'impression instantanée ou « snapshots » permet de récupérer les données perdues. Ces « snapshots » sont effectués toutes les heures, et gardés 24h. De plus, un « snapshot » est effectué une fois par nuit et conservé 1 an. De plus, une synchronisation sur un serveur hébergé sur le mésocentre de l'Université Clermont-Auvergne, est effectuée tous les soirs. Ce serveur est synchronisé avec une brique distante AgroDataring située dans le Datacenter INRAE de Toulouse.

Sur la partie universitaire de la PFEM, un script recopie tous les soirs les données produites par les matériels d'acquisition sur un serveur de stockage. Le serveur est sauvegardé tous les soirs sur disque au mésocentre de l'Université Clermont-Auvergne. La stratégie de sauvegarde mise en œuvre est basée sur une rétention de 3 mois.

Concernant le système d'information de la PFEM, l'ensemble des bases de connaissances et fichiers utilisées ou produits par nos outils de gestion sont sauvegardées sur un serveur dédié avec une rétention de 5 jours. L'activité de stockage et sauvegarde des données conduite au sein de la PFEM est décrite dans le système qualité de l'infrastructure, notamment dans sa procédure « Gestion des données et de l'architecture informatique de la PFEM ».

Quelles sont les mesures de sécurité mises en place lors des étapes de transfert des données ?

- **Cas des données analytiques.**

Les données brutes générées sont stockées sur un serveur de stockage local du centre INRAE de Theix. Seuls les personnels et les responsables de la PFEM ont un accès à ces données brutes, soumis à une authentification personnelle (authentification LDAP INRAE). Ces droits d'accès sont accordés par les responsables de la PFEM au vu des fonctions des agents, et gérés par les administrateurs système de la PFEM. La sauvegarde de ces données brutes est effectuée sur un autre serveur sur un site géographique différent de celui du stockage (le mésocentre de l'Université Clermont-Auvergne à Aubière). Seuls les personnels, administrateurs système de la PFEM, ont accès à ces sauvegardes par un accès sécurisé. En cas d'incidents, les données sauvegardées ou archivées peuvent être récupérées via des outils internes par les administrateurs système de la PFEM. A la fin du projet, les données sont archivées sur le datacenter INRAE. Seuls les administrateurs système de la PFEM ont accès à ces archives (accès SSH). La gestion des accès et la sécurité des données est décrite dans la procédure P-DOC-06 (Gestion des données et de l'architecture informatique de la PFEM).

Les données prétraitées et traitées ainsi que les fichiers de métadonnées sont générées par les agents de la PFEM, grâce à des applications en ligne et logiciels disponibles sur un serveur de la PFEM. En cas de collaboration, un partage peut être mis en place, en interne via authentification (LDAP INRAE) ou en externe via un service de partage institutionnel avec contrôle des accès. Ces droits d'accès sont accordés par les responsables de la PFEM au vu des fonctions des agents, et gérés par les administrateurs système de la PFEM. La sauvegarde de ces données prétraitées et traitées est effectuée sur un autre serveur sur un site géographique différent de celui du stockage (Mesocentre UCA à Aubière). Seuls les administrateurs système de la PFEM ont accès à ces sauvegardes (accès SSH). En cas d'incidents, les données sauvegardées ou archivées peuvent être récupérées via les procédures en vigueur et les outils internes par les administrateurs système de la PFEM.

L'intégrité des fichiers est vérifiée lors des transferts entre serveurs par utilisation de leur empreinte SHA256 de hachage.

- **Cas des données informatiques**

L'ensemble des données informatiques (logs, sauvegardes, bases de données, …) sont accessibles aux seuls

administrateurs du SI de la PFEM.

Quelle est la volumétrie actuelle et prévisionnelle ?

La PFEM stocke et archive actuellement 80 To de données issus des projets déroulés de 2010 à 2024.

La production actuelle est de 1,5 To par an avec une augmentation plannifiée en lien avec l'arrivée de trois nouveaux instrument de plus grande précision en 2024.

L'entité hébergeant physiquement les données a-t-elle une politique de sécurité de l'information et a-t-elle un plan d'assurance sécurité ?

En tant que plate-forme portée par INRAE et par l'Université Clermont-Auvergne, la PFEM veille à respecter d'une part la Charte nationale de déontologie des métiers de la recherche dont INRAE est signataire ainsi que la Charte INRAE de déontologie, d'intégrité scientifique et d'éthique des projets de recherche, et d'autre part le décret relatif au respect des exigences de l'intégrité scientifique par les établissements publics contribuant au service public de la recherche qui promeut des exigences « FAIR » et science ouverte mais aussi de déontologie encadrées par la loi.

- **Respect des dispositions légales pour la production de données**

Dans le cas des projets de recherche externes, la production des données analytiques est dépendante des échantillons biologiques analysés. En cas de partenariat, le porteur du projet est garant de la conformité de son projet vis-à-vis des questions légales, éthiques et déontologiques. En fonction de son domaine de recherche, le partenaire fait évaluer en amont son protocole de recherche par les comités adéquats (e.g. comités d'éthiques nationaux) qui autorisent la tenue du projet. Lorsque le projet de recherche ne nécessite pas d'évaluation par une commission externe, c'est le porteur du projet ayant généré les échantillons fournis qui est garant de l'adéquation des analyses prévues avec les questions d'éthique et de déontologie.
Dans le cas des projets internes, une attention particulière est portée pour que les objectifs technologiques et scientifiques ainsi que la nature des échantillons analysés ne nécessitent pas une revue des pratiques par un comité externe. Les thématiques abordées ne présentent pas de risque notable vis-à-vis de questions éthiques et de déontologie, qu'elles soient d'ordre analytiques (exemple : qualité du signal ou annotation), en lien avec le traitement de données (exemple : gestion d'effets analytiques dans un jeu de données) ou la construction de systèmes d'informations (exemple : conception de base de connaissance).
Si la PFEM est amenée à produire des données analytiques internes, l'utilisation d'échantillons standardisés commerciaux (exemple : standard, plasma NIST) sera privilégié. Sinon la production de données analytiques internes peut se faire à partir d'échantillons biologiques avec l'accord du porteur du projet ayant généré ces échantillons qui sera alors le garant de l'adéquation des analyses prévues avec les questions d'éthique et de déontologie.
Dans le cas des algorithmes et des codes sources, leur production s'inscrivant dans les missions de la PFEM se fait en conformité avec la législation française et internationale, des recommandations du Cigref et du Syntec ou de l'IFIP et ne présente pas de préoccupation éthique ou déontologique spécifique à l'activité de la plate-forme.

- **Autres questions éthiques et déontologiques**

La PFEM a une obligation de mise en œuvre des moyens prévus dans le cadre de ses missions mais pas de résultats. En cas de dysfonctionnement avéré dans les processus de production ou de réutilisation de données, la plate-forme informe ses partenaires de la survenue de l'incident et met en œuvre les moyens à sa disposition pour permettre une rectification. A noter que le système de management de la qualité de la PFEM prévoit la gestion des anomalies/incidents.
La PFEM promeut la science ouverte, qui par ailleurs est une des orientations politiques d'INRAE et de l'UCA, ainsi qu'un des axes promus nationalement et au niveau européen pour la recherche. En particulier, elle porte une attention particulière à ce que la production et la ré-utilisation des données s'inscrivent au mieux dans les principes FAIR, et demande à ses partenaires qu'ils soient en mesure de fournir un argumentaire clair en cas de choix de fermeture des données. Pour rappel du cadre légal, un guide est disponible pour aiguiller les chercheurs vis-à-vis de l'obligation de partage ou non des données.

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 ?

Question sans réponse.

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

- **Responsabilités de gestion des données**

La production des données, la production des métadonnées d'analyse et la vérification de la qualité des données instrumentales sont assurées par les analystes chimistes de la plateforme sous la responsabilité des responsables de la PFEM.
La production des données, la production des métadonnées et la vérification de la qualité des données issues du traitement sont assurés par les statisticien.ne.s et bioinformaticien.nes de la plateforme sous la responsabilité des responsables de la PFEM. Le stockage et sauvegarde, archivage et partage des données produites sont assurés par les administrateurs systèmes de la plateforme sous la responsabilité des responsables de la PFEM.
L'infrastructure PFEM s'appuie sur une équipe en charge de la mise en œuvre et de l'évolution au fil du projet du plan décrit dans ce document avec notamment l'administrateur des données de l'infrastructure et un réseau de référents opérationnels « données » locaux répartis par domaine (production de données, traitement de données, gestion des données et management).
La PFEM s'engage à fournir les informations (décrites dans le présent document) nécessaires à l'implémentation des futurs plans de gestion de données de ces projets ainsi que les outils ou les stratégies nécessaires à l'ouverture des données en lien avec les études de métabolomique.

Les responsabilités légales sont définies dans le contrat de recherche. En cas d'absence, INRAE est alors responsable des données produites. Les termes des présentes conditions d'utilisation peuvent être amendés à tout moment, sans préavis, en fonction des modifications apportées à la Plateforme, de l'évolution de la législation ou pour tout autre motif jugé nécessaire.

- **Ressources de gestion des données**

La PFEM dispose des expertises et des personnels permanents nécessaires à la mise en place de sa politique de gestion de données. Elle incite néanmoins les porteurs des projets scientifiques à inclure des ressources contractuelles chargé de cette gestion dès l'élaboration de leurs propositions de projets.

Concernant l'infrastructure informatique, la PFEM s'appuie sur les investissements réalisés (budget + ressources humaines) ainsi que de ressource de projets nationaux tels que le projet de stockage collaboratif AgroDataRing soutenu et impliquant l'infrastructure nationale MetaboHUB avec notamment :

- Un budget annuel dédié au fonctionnement et l'évolution de l'e-Infrastructure.
- Des personnels en charge de l'e-Infrastructure notamment en administration système et en développement informatique pour la maintenance des outils

Les agents de la structure ont-ils bénéficié d'une sensibilisation aux bonnes pratiques d'hygiène numérique ?

- Oui
- Oui

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 PFEM s'engage à assurer une sécurisation des données (accès, sauvegarde, ...) à l'état de l'art ainsi qu'à leur conservation sur des pas de temps alignés avec la législation (Exemple : 10 ans de conservation des données humaines).

L'ensemble des données produites sont conservées sur le moyen terme en étant disponible en ligne pour les membres du projet et ses personnels et sur le long terme (10 ans) sur un volume à accès tres restreint (adminstrateurs PFEM).
A noter que la PFEM n’assure pas et ne propose pas de service d’archivage long terme de données (supérieur à 10 ans).

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 ?

A noter que la PFEM n’assure pas et ne propose pas de service d’archivage long terme de données (supérieur à 10 ans).

Quelle est la durée de conservation des données ?

Question sans réponse.

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

Question sans réponse.