DMP du projet "Plan de gestion de données de la plateforme métabolisme métabolome version 5.3"

Valérie Cantonny, Bertrand Gakière, Françoise Gilard

▶ **To cite this version:**

Valérie Cantonny, Bertrand Gakière, Françoise Gilard. DMP du projet "Plan de gestion de données de la plateforme métabolisme métabolome version 5.3". Inist-CNRS. 2018. hal-04437422

HAL Id: hal-04437422
https://hal.inrae.fr/hal-04437422v1

Submitted on 4 Feb 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

DMP du projet "Plan de gestion de données de la plateforme métabolisme métabolome version 5.3"

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

Renseignements sur le plan

Titre du plan	DMP du projet "Plan de gestion de données de la plateforme métabolisme métabolome version 5.3"
Langue	fra
Date de création	2019-11-18
Date de dernière modification	2020-05-15
Identifiant	4234

Renseignements sur le projet

Titre du projet	Plan de gestion de données de la plateforme métabolisme métabolome version 5.3
Résumé	La plateforme produit des données qualitatives et quantitatives en métabolomique et en isotopie. Elle produit également des résultats interprétés sous forme de statistiques. Les données sont collectées sous différentes formes, et sous différents formats. Elles sont stockées, sauvegardées, archivées et reliées entre elles, afin d'en assurer la traçabilité, et leur réutilisation ultérieure à différents niveaux de traitement.

Produits de recherche :

1. Plan de gestion des données de la plateforme métabolisme métabolome (Jeu de données)

Contributeurs

Nom	Affiliation	Rôles
Bertrand Gakière - https://orcid.org/0000-0003-0285-5160		• Coordinateur du projet • Personne contact pour les données
Cantonny		• Responsable du plan

DMP du projet "Plan de gestion de données de la plateforme métabolisme métabolome version 5.3"

Informations sur la structure

Nom de la structure

Plateforme Métabolisme Métabolome de l'IPS2: PMM

Type de structure

- Plateforme, plateau technique
- Plateforme, plateau technique

La plateforme est hébergée par l'Institut des Plantes de Paris Saclay qui est membre de l'Ecole Universitaire de Recherche Saclay Plant Sciences Graduate School of Research (EUR SPS-GSR). Elle est rattachée à l'équipe de recherche Métaboactions: Signalisation, régulation et interactions métaboliques.

Au sein de l'IPS2 la plateforme s'intègre dans une collaboration étroite inter-plateforme: le réseau SPOMICS (Saclay Plant Omics), avec les plateformes POPS (Transcriptomique) et InterATOME (Plateforme d'étude des interactions protéines-protéines) . Les trois plateformes gèrent et intègrent des données moléculaires à haut débit dans le domaine de la biologie végétale.

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

Question sans réponse.

Responsabilités dans la structure

Nom, Prénom	Courriel	Rôle
Françoise Gilard	francoise.gilard@u-psud.fr	Responsable de la plateforme
Gakière Bertrand	bertrand.gakiere@u-psud.fr	Responsable scientifique
Caroline Mauve	caroline.mauve@u-psud.fr	Référente qualité, responsable des techniques HPLC, RMN, CI, LC-IRMS et GC-C-IRMS
Florence Guérard	Florence.guerard@u-psud.fr	Rresponsable des techniques GC MS, UPLC, LC-MS-MS et MS-MS
Valérie Cantonny	valerie.cantonny@u-psud.fr	Responsable de la conservation à long terme et de la traçabilité des données

Etablissement(s) tutelle(s)

INRA, CNRS, Université Paris-sud, Université Paris Diderot, Labex SPS

Département de rattachement Inra

- BAP
- BAP

Biologie et amélioration des plantes

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

INRAE, CNRS, IPS2, UPSUD, utilisateurs de la plateforme.

Informations sur le plan de gestion

DOI (version publiée du plan de gestion)

10.15454/RVWYM0

Historique des versions

Date	n° de version	Status	Auteur	Affiliation de l'auteur *(se reporter à l'annuaire Inra)*	Validé par	Validé le
				IDF Versailles Grignon		
24/12/2019	5.3	Rajout des logiciels propriétaires et libre		IDF Versailles Grignon		
24/12/2019	5.2	Modification des deux schémas	Valérie Cantonny	IDF Versailles Grignon		
24/12/2019	5.1	Relecture, mise en forme du fichier word	Valérie Cantonny	IDF Versailles Grignon		
23/12/2019	5.0	Rajout des formats des données, réagencement des paragraphes	Valérie Cantonny	IDF Versailles Grignon		
20/11/2019	4.1	Rajout de la partie sur la conservation des données	Valérie Cantonny	IDF Versailles Grignon		
18/11/2019	4.0	en cours de rédaction		IDF Versailles Grignon		
13/11/2019	3.0	Template:INRA__Trame_Structure : gardé pour model		IDF Versailles Grignon		
06/11/2019	2.0	annulé: Template Horizon_2020_FAIR_DMP_franais		IDF Versailles Grignon		
30/10/2019	1.0	annulé: Template: CIRADTEMPLATE	Valérie Cantonny	IDF Versailles Grignon		

Présentation générale des données

Mode d'obtention des données

- Données générées par la structure
- Données générées par la structure

Les données générées par la plateforme proviennent de l'analyse d'échantillons, principalement de plantes, par différents équipements tels que des spectromètres ou des Chromatographes.

Origine

- Analyse
- Expérimentation
- Analyse
- Expérimentation

Les métadonnées et données de la plateforme proviennent de trois origines :

- Les conditions de culture, la taxonomie des échantillons
- La préparation des échantillons par la plateforme afin qu'ils puissent être analysés par les équipements de la plateforme. Elles sont décrites par les protocoles de préparation
- Les analyses qui sont réalisées par couplage de différents équipements.

Descriptif des différents équipements de la plateforme :

- Couplage Chromatographie en Phase Gazeuse avec un spectromètre de masse: Gas Chromatography Mass Spectrometry (GC MS) permet d'identifier environs 700 métabolites extraits d'un échantillon
- Chromatographie liquide ou ionique: Liquid Chromatography (LC) ou pour le dosage de métabolites spécifiques: acides aminés (sauf la proline), les sucres (Saccharose, Glucose et Fructose), les cations (Li+,Na+,NH4+,K+,Mg+,Ca+) et les anions (F-, Cl-,NO2-,Br-,NO3-,P-,SO3-)
- Chromatographie en phase liquide couplée à un spectromètre de masse: Liquid Chromatography Mass Spectrometry (LC-MS et LC-MS-MS), pour la quantification de 25 cofacteurs-nucléotides préalablement extraits d'une matrice, ainsi que des mycotoxines.
- Spectrométrie de masse à ratio isotopique: Isotope-Ratio Mass Spectrometry IRMS). La Plateforme possède deux spectromètres de masse isotopique (IRMS) et plusieurs systèmes couplés permettant d'obtenir un gaz pur à partir d'échantillons solides, liquides et gazeux. Les spectromètres de masse isotopiques mesurent le rapport isotopique (13C/12C, 15N/14N, 18O/16O ou encore D/H) du gaz pur (CO2, N2, O22) issu d'échantillons solides ou liquides et le comparent à celui d'un gaz de référence (étalonné par rapport au standard international).
- Résonance Magnétique Nucléaire: Les analyses RMN permettent l'acquisition d'expériences 1 Dimension : H, 13C, 15N et 31P et/ou 2 Dimensions homo ou hétéro nucléaire (ex : HSQC, COSY….).Ces acquisitions peuvent permettent la quantification absolue dans la matrice (ex : Activité enzymatique), l'identification structurale ou encore de pouvoir de suivre un enrichissement isotopique dans les positions intramoléculaires.

Ref: http://ips2.u-psud.fr/fr/plateformes/spomics-interatome-metabolome-transcriptome/pmm-plateforme-metabolisme-metabolome/analyses-disponibles.html

---

Type de données

- Dataset
- Text
- Workflow
- Dataset
- Text
- Workflow

Les métadonnées de type **texte** concernent les conditions de culture, la description des échantillons, la préparation des échantillons. Elles sont décrites par l'application ARMOISE Analytical Repository for MetabOlomics and stable ISotope Entries, qui a été programmée en php/postgresQL par madame Cantonny.

On distingue trois types de données, obtenues lors d'un projet d'analyse

- les données brutes, elles proviennent directement de l'équipement scientifique, souvent elles sont dans un format propriétaire, sous forme de **Dataset**
- Les données traitées ont été recalculées à partir du logiciel du fournisseur, d'un logiciel libre ou d'un simple tableur, ce sont des données de type **texte** ou des **dataset**.
- Les données interprétées sont sont l'ensemble des fichiers servant à la préparation du compte rendu de résultats. Elles comprennent : des analyses statistiques, des graphes, des diagrammes, des données de type **texte** et des images.

Ces trois types de données décrivent un **workflow** qui est explicité dans le protocole qualité de la plateforme, nommé cycle des données: "Cycle_des_donnees_PCS1_PO9055_2018_12_19_ver7".

Nature des données

Les données se trouvent sous ces différentes formes:

- Chromatogrammes
- Fichiers binaires crées lors des analyses
- Fichier textes
- Histogrammes
- Images

Format des données

Les formats de données générées par la plateforme sont très diverses suivant les spectromètres utilisés et les niveaux de calculs de ces données :

Technique	DB	DT	DI
HPLC	.EXP .INF .CKS .dat	.xls	.xls
FlashEA	.DAT	.xls	.doc .xls
EaIrmsPyro	.raw .xls .pdf .xps	.xls	.xlsx .pdf
GccCIrms	.pro .xls .doc .bmp .pdf	.raw .xls .pdf .doc	.xls
GcMs	.D .xml .txt .ini .csv .M .ms .m .bin	.Cal .ini .m .xml	.xlsx .txt .anl .xlsm .JNB .pdf .anl .docx .bmp
IC	.cmb		.docx .xlsx
LcMsMs	.d .ini .xml .m .cg .sd .sp .bin .xsd .cd .stg .clc	.d .ini .xml .m .cg .sd .sp .bin .xsd .cd .stg .clc	.xlsx .docx .pdf
RMN	.txt .info .par .xml .png	.pdf .txt .info .par .xml .png	.xlsx .docx

---

Périmètre thématique des données

- Omics
- Chemistry and chemical engineering
- Omics
- Chemistry and chemical engineering

Le rôle de la Plateforme consiste en l'établissement de profils métaboliques (GC-MS, LC-MS), l'identification, la quantification des métabolites d'intérêt (RMN, LC), la détermination éventuelle de leur composition isotopique (IRMS et GC-C-IRMS) et le suivi de marquage (RMN).
Réf: http://ips2.u-psud.fr/fr/plateformes/spomics-interatome-metabolome-transcriptome/pmm-plateforme-metabolisme-metabolome/analyses-disponibles.html

---

Droits de propriété intellectuelle

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

Lorsqu'un chercheur fait une demande d'analyse, il signe un accord de paternité : "FicheAccordPaternite_PCRE1_EN6010_v2" dans lequel il s'engage "à associer la Plateforme à toute production scientifique utilisant des données produites par la Plateforme, avec pour co-auteurs le ou les ingénieurs en charge du projet d'analyse ainsi que le Directeur Scientifique de la Plateforme. "

La plateforme s'engage à ne pas divulguer les données avant publication des résultats ou accord du chercheur à l'origine de la donnée d'analyse.

La plateforme est responsable des données qu'elle a produites, mais ces données « *appartiennent au client* ***qui cède tacitement*** *un droit d'utilisation à la Plateforme pour l'extraction des résultats finaux et la rédaction du compte-rendu.* » (Manuel qualité de la plateforme p.16)

Confidentialité

Identification des jeux de données contenant des données confidentielles

Les données sont reliées entre elles par l'application web/base de données ARMOISE. La base de données permet l'identification unique des données brutes, traitées et interprétées et des projets scientifiques qui ont fait l'objet d'une demande d'analyse. Les données de cette application représentent **l'architecture de nos données privées.**

En cas de restriction à l'accès aux données, seul le chercheur et les analystes de la plateforme ont accès aux données. Le chercheur peut télécharger ses données interprétées au format Excel depuis l'application ARMOISE. Les données sont stockées, sauvegardées et archivées sur les serveurs de l'université Paris sud. Elles sont identifiables via l'application ARMOISE. Leur mise à disposition se fait à la demande.

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

La sauvegarde et l'archivage des données de la plateforme sont sous la responsabilité du service informatique de l'IPS2. Nous devons respecter la politique de sécurité de l'institut. Ainsi chaque utilisateur doit signer la charte du service informatique et la charte qualité de la plateforme "Charte_de_la_PlateformePCS5_PO7029_11032015_3".

Les PC de pilotage des spectromètres sont uniquement reliés au réseau de l'IPS2. Ils ne possèdent pas de connexion internet. La plateforme possède également des PC de traitement de données pour générer les données traitées et interprétées. Ces PC ne sont pas reliés au réseau interne de l'IPS2, ils ont la possibilité de se connecter au serveur de l'IPS2 pour récupérer les données brutes.

Concernant l'application ARMOISE, une déclaration à la CNIL (numéro 2-14026) avait été faite en 2014, car la base de données contient les noms des utilisateurs de la plateforme.

L'application ARMOISE possède un système d'identification par connexion internet. Les données ne sont pas publiques.

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 ?

Concernant les techniques LC-MS, GC-MS et RMN, des analyses de second niveau peuvent être réalisées par l'application Workflow4metabolomics.

Celle-ci nécessite une authentification lors de son utilisation.

Partage des données

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

Il n'y a pas d'obligation de partage des données sur la plateforme, mais la plateforme encourage ses clients à effectuer cette démarche.

Quelles sont les réutilisations potentielles de ces données ?

La réutilisation des données permet la comparaison des jeux de données issues d'expériences et de plateforme différentes. Les données peuvent être analysées par d'autres logiciels pour confirmer ou vérifier un résultat.

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

La plupart des données brutes sont lisibles par le logiciel du PC de pilotage de l'équipement qui les a produites :

Technique	Logiciel
HPLC	Empower
Flash EA	Eager300 for EA 112
EA IRMS Pyrocube	Ion Vantage 1.7.3.0, Vario Pyrocube 3.2.3
GCC C IRMS	Ion Vantage, Agilent Chemstation B04.03
GC MS	Suite Agilent MassHunter, MS Search V2.2
IC	Chromeleon 6.8
LC MS MS	Suite Agilent MassHunter
RMN	Topspin 3.2

Les logiciels libres utilisés sur la plateforme sont Amdis pour visualiser les spectres de masse obtenus en GC MS, MEV v4.9.0 pour les statistiques

---

Comment les données seront-elles partagées ?

A chaque préparation de publication, les données d'analyse sont chargées dans l'application MetaboLights sous réserve de l'accord du chercheur qui a fait la demande d'analyse. Elles sont validées par les administrateurs de l'application.
Par exemple, le projet de collaboration avec Christophe Bailly est partagé sur MetaboLights à cette adresse : https://www.ebi.ac.uk/metabolights/MTBLS662 :
"*Regulatory actors and alternative routes for Arabidopsis seed germination are revealed using a pathway-based analysis of transcriptomic datasets*"
Actuellement, nous étudions la possibilité de les déposer dans le Dataverse de l'INRAE.

Avec qui ?

- Tous (open acces)
- Tous (open acces)

Lorsque les données sont partagées, elles sont visibles et téléchargeables par tout le monde.

Sous quelle licence ?

- Autre (à préciser dans la zone d'Informations supplémentaires
- Autre (à préciser dans la zone d'Informations supplémentaires

Je n'ai pas trouvé de licences pour les données chargées dans l'application, j'ai envoyé un mail pour demander si une licence s'applique automatiquement aux données que nous chargeons dans l'application Metabolights.
L'application ARMOISE est sous licence GNU General Public Licence

---

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

Lorsqu'un chercheur souhaite faire une demande d'analyse, il s'entretient avec les deux coordinateurs de la plateforme sur la faisabilité de son projet, ce qui lui permet également d'affiner son choix quant à l'utilisation de la technique d'analyse la plus appropriée. Il remplit une demande d'analyse comportant la thématique du projet, le nombre d'échantillons à analyser et le tarif. Il effectue une demande d'analyse dans l'application ARMOISE de la

plateforme. Un devis et un accord de collaboration scientifique avec copaternité des résultats lui sont proposés. Le numéro de devis est également saisi dans l'application ARMOISE et est relié à la demande d'analyse.

Les échantillons sont préparés par le client selon le protocole qualité adapté à la technique d'analyse. Les analystes réalisent les manipulations nécessaires au bon déroulement de l'analyse par l'équipement choisi: RMN, GCMS...

Les données traitées sont recalculées via un tableur, excepté pour la GC MS, pour laquelle nous utilisons le logiciel libre AMDIS.

Les données interprétées sont également traitées via un tableur sur les PC de traitement de la plateforme. Les statistiques sont réalisées par le logiciel Sigma plot pour l'instant. Nous souhaitons utiliser R dans les 3 années à venir.

Lorsqu'une analyse est terminée, l'analyste responsable du projet envoie les résultats par mail au chercheur. Il met l'informaticienne en copie afin qu'elle puisse relier les données d'analyses entres elles dans l'application armoise

---

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 ?

Métadonnées	**Origine, mode de production des métadonnées (ex : saisie manuelle, annotation atomatique...)**	**Standard, Vocabulaires associés**
Taxonomie des plantes	Saisie via l'interface web Metabolight	Vocabulaires contrôlés pour différentes métadonnées (sous-ensemble d'OBI Ontology et NCBI Taxonomy for Organisms)
Description du projet, du demandeur du projet et description administrative	Saisie dans la base de données de l'application ARMOISE et dans l'application Metabolight	Base de données relationnelle sous postgreSQL, outils ISAcreator pour l'application Metabolight
Description des échantillons et de leur préparation.	Base de données de l'application ARMOISE, saisie via des fichiers excels dans l'application Metabolight.	Protocoles qualité d'injection de dérivation et de préparation des échantillons
Description des métabolites trouvés	Dans l'application Metabolight	CHEBI, formules, Smiles, InCHi

---

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

Les protocoles qualité d'injection de dérivation et de préparation des échantillons constituent des documentations complémentaires à la préparation des échantillons.
Le cahier des charges de l'application armoise, contient un Model Conceptuel de Données qui décrit les métadonnées.

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

Chaque type de données répond à une nomenclature qui a été déterminé dans le protocole qualité du cycle des données, "Cycle_des_donnees_PCS1_PO9055_2018_12_19_ver7":

- Données brutes et traitées: **annéemois_NomInterlocuteur_Autre,** avec
  - **annéemois:** Devis de démarrage des analyses (acquisition), le jour est facultatif
  - **NomInterlocuteur:** La personne ou l'organisme qui a demandé les analyses ou ses initiales
  - **Autre:** Le numéro de l'analyse si il y a plusieurs analyses pour le même interlocuteur et sur le même devis pour le même mois, ou des précisions particulières
- Données interprétées: **technique_annéemois_n°Devis_NomInterlocuteur_Autre**
  - **technique**: GCMS
  - **annéemois:** Devis de démarrage des analyses
  - **n°Devis:**
  - **NomInterlocuteur:** La personne ou l'organisme qui a demandé les analyses ou ses initiales
  - **Autre:**

La conservation des données est garantie pour 12 ans.

Quelle est le processus de contrôle qualité des données ?

Le fonctionnement de la plateforme est régit par le système de management de la qualité qui a été mis en place dès 2010. Celui-ci est conforme aux exigences du référentiel qualité FD-X-50-550 de la norme **ISO 9001** version 2000 (2008)

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)

Les données brutes sont lues et processées par le logiciel du PC de pilotage des équipements d'analyse.
Elles sont stockées sur les PC de pilotage, puis transférées sur le serveur de l'IPS2.
Les données traitées sont récupérées depuis le serveur de l'IPS2 vers les PC de traitement de la plateforme pour être recalculées. Lorsque les calculs sont terminés les données sont sauvegardées dans leur répertoire respectif. Les données sauvegardées sont conservées pendant 6 ans sur le serveur de l'université. Elles sont ensuite archivées pendant 6 ans.

Quelle est la volumétrie actuelle et prévisionnelle ?

La plateforme produit environ 1To de données par an.

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...

La politique de sauvegarde de l'université Paris Sud qui héberge les données de la plateforme consiste à sauvegarder 2 snapshots (cliché des données à un instant donné) par jour et un répliquât du filer DELL, sur lequel se trouvent les données d'analyses, d'interprétations et de description des échantillons.

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 ?

Il n'y a pas de droit d'accès aux données avant leur publication, excepté par le chercheur, son équipe, et le personnel de la plateforme

---

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

La base de données de la plateforme et le code PHP de l'application ARMOISE sont hébergées sur un serveur de l'université Paris Sud. La sécurité et la confidentialité des données est assurée par l'obligation de se connecter via l'interface Web. Celle-ci a été programmée dans le Framework Jelix, au format MVC (Modèle, Vue, Controlleur), ce qui accroit la sécurité et l'intégrité du code.

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 totalité des données sont sauvegardées pendant 6 ans puis archivées pendant 6 ans puis détruites

---

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 ?

Les données sont archivées sur un serveur de l'université pour le premier niveau et dans le Datacenter d'un site distant pour le deuxième niveau.

Quelle est la durée de conservation des données ?

Les données sont conservées pendant 12 ans

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

Question sans réponse.