DMP du projet "DMP plateforme Metatoul"

Plan de gestion de données créé à l'aide de DMP OPIDoR, basé sur le modèle "ANR - Modèle de PGD (français)" fourni par Agence nationale de la recherche (ANR).

Renseignements sur le plan

Titre du plan	DMP du projet "DMP plateforme Metatoul"
Version	Version intermédiaire
Domaines de recherche (selon classification de l'OCDE)	Biological sciences (Natural sciences), Agricultural biotechnology, Health biotechnology, Industrial biotechnology, Biological sciences (Natural sciences)
Langue	fra
Date de création	2022-06-10
Date de dernière modification	2023-11-21
Identifiant	MetaToul_2023
Type d'identifiant	Identifiant local
Licence	**Nom**: Creative Commons Attribution 4.0 International **URL**: http://spdx.org/licenses/CC-BY-4.0.json

Renseignements sur le projet

Titre du projet	DMP plateforme Metatoul
Acronyme	Metatoul
Résumé	La plateforme Metatoul propose son expertise dans le domaine de l'analyse et de la compréhension du métabolisme. Elle regroupe des compétences et des technologies de pointe en résonance magnétique nucléaire, spectrométrie de masse, robotique, (bio)-informatique, biostatistiques, biochimie, qu'elle met à disposition des communautés scientifiques académique et industrielles. Ses sites spécialisés développent et proposent des outils et concepts pour l'analyse du métabolisme à l'échelle d'un système biologique (cellule, tissu, organe, organisme) : • Metatoul-Agromix : analyses qualitatives et quantitatives de métabolites de plantes et microorganismes associés par des approches ciblées ou globales. • Metatoul-Axiom : prises d'empreintes métabolomiques sans a priori (RMN, MS), analyses qualitatives et quantitatives de xénobiotiques et métabolites (exposome), identification de métabolites d'intérêt toxicologique, imagerie par spectrométrie de masse de métabolites, analyses statistiques. • Metatoul-FluxoMet : analyse fonctionnelle des réseaux métaboliques par des approches d'analyse quantitative, profilage isotopique et fluxomique (RMN, MS, outils bioinformatiques) • Metatoul-FluxoVivo : profilage isotopique et fluxomique sur organismes entiers (modèles animaux, Homme) • Metatoul-Lipidomique : analyses qualitatives et quantitatives de différentes familles lipidiques par des approches ciblées ou globales. • Metatoul-MetExplore : bioinformatique pour l'analyse des données dans le contexte des réseaux métaboliques

Produits de recherche :

1. Jeu de données MS, RMN, brutes (raw) et retraités (Jeu de données)
2. Developpement logiciels (Logiciel)

Contributeurs

Nom	Affiliation	Rôles
Bellvert Floriant	CNRS	• Coordinateur du projet • Personne contact pour les données (Dataset)
Marti Guillaume		• Responsable du plan de gestion de données
RIFA Etienne		• Personne contact pour les données (Software)

DMP du projet "DMP plateforme Metatoul"

1. Description des données et collecte ou réutilisation de données existantes

Jeu de données MS, RMN, brutes (raw) et retraités

1a. Comment de nouvelles données seront-elles recueillies ou produites et/ou comment des données préexistantes seront-elles réutilisées ?

Nous distinguons deux étapes de génération des données, et deux type de réutilisation des données existantes:

1) Données brutes générées par les mesures instrumentales sur les échantillons (EJA nous sommes d'avis sur AXIOM de ne pas se restreindre à la notion d'échantillon biologique car il nous arrive d'analyser des échantillons environnementaux)

2) Données générées par le traitement des données instrumentales

3) Réutilisation de métadonnées liées aux informations des échantillons

4) Réutilisation de données internes ou externe à la plateforme

1) Données brutes d'origines instrumentales

Il s'agit des données brutes (données spectrales MS et RMN) acquises sur des instruments d'analyses. Ces mesures physico-chimiques sont réalisées sur des échantillons biologiques. Les données relatives à un échantillon analysé vont être collectées sous un fichier (ou un ensemble de fichiers) au format défini par le constructeur de l'instrument (format propriétaire). Une fois la donnée instrumentale obtenue, des solutions logicielles existent pour la convertir si nécessaire en un format standardisé ouvert (exemple : nmrML ou mzML).

2) Données générées par traitement de données

Les données d'origine instrumentale nécessitent des étapes de traitement pour être interprétables. Ceci couvre un ensemble vaste de processus qui sont sollicités en fonction de l'appareil utilisé pour générer la donnée initiale mais aussi des choix de stratégies de traitements en lien avec la question scientifique relié à l'échantillonnage. Ces processus génèrent des données intermédiaires et finalisées, qui nécessitent parfois la prise en compte de données préexistantes tiers caractérisant les échantillons analysés (exemple : détails de protocoles cliniques).

Quelle que soit la nature de la donnée produite, le processus de production de cette donnée génère un ensemble d'informations caractérisantes (ex : date de génération de la donnée, nature du processus la générant). Ces informations peuvent être formalisées directement par le processus de production (exemple : log d'outils Galaxy), ou bien réunies via une procédure indépendante (ex : traçabilité décrite dans les systèmes qualité des plateformes de l'infrastructure).

La plateforme **Metatoul** pourra également **réutiliser des données existantes**, notamment :

3) Données liées directement à des échantillons analysés (données de demandeurs/partenaires)

Chaque projet analytique est discuté en amont de la génération des données, et les informations minimales descriptives des échantillons allant être analysés sont discutées avec le demandeur. C'est le demandeur qui fournit ensuite les données nécessaires ainsi que les conditions de réutilisation spécifiques de ces données. La propriété intellectuelle des données produites avec le concours de ces données pré-existantes est définie en amont de leur production en accord entre le demandeur et le ou les membres de **Metatoul** et cela à travers leur(s) tutelle(s) d'adossement.

4) Données précédemment produites par les instruments de la plateforme Metatoul

Les données précédemment produites par les membres de **Metatoul** peuvent être réutilisées par ceux-ci en accord avec les conditions de ré-utilisation définies à la génération initiale de la donnée. En particulier, les données d'origine instrumentale peuvent être réutilisées en interne des laboratoires à des fins de développement et d'amélioration de processus.

Données publiques issues de la littérature

Les données qui ont été publiées ou mises à disposition de façon ouverte sont réutilisables en accord avec la licence associée à la publication des données concernées.

1b. Quelles données (types, formats et volumes par ex.) seront collectées ou produites ?

Une description détaillée des données qui seront collectées ou générées dans ce cadre ainsi que des données qui seront réutilisées dans le projet se trouve dans le tableau d'inventaire des types de données (DIT) ci dessous qui sera mise à jour en fonction des évolutions notamment en terme d'instrumentations de la plateforme:

Catégorie	Origine de la donnée	Format	Type de donnée	Estimation/volumétrie	commentaires
Données brutes MS	Spectromètres de masse Thermo, Waters	*.raw	Propriétaire/spécifique du fournisseur logiciel	14To/an	Conversion au format ouvert
Données brutes MS	Spectromètres de masse Bruker	*.tdf	Propriétaire/spécifique du fournisseur logiciel	6To/an	Conversion au format ouvert
RMN	MicroNMR probe	*.d	Propriétaire/spécifique du fournisseur logiciel	2To/an	Conversion au format ouvert
Données converties	LCMS	mzML	Format ouvert de données spectrales avec un minimum de métadonnées	20To/an	Format ouvert, standard du domaine
Données converties	RMN	NMRML	Format ouvert de données spectrales avec un minimum de métadonnées	2To/an	Format ouvert, standard du domaine
Données traitées	pipeline de traitement	.csv, tsv, txt...	données tabulées	1Go/an	Traitements des signatures spectrales
Données traitées	Outils de bioinformatique	xlsx, svg, pdf, tsv, yaml, kvh, csv, png, jpeg, attrs	Données tabulées, images, données de configuration, mappings	10Go/an	Traitement de données par outils bioinformatiques développés in-house
LCMS	LTQ-Orbitrap QExactive+	.raw	Propriétaire/spécifique du fournisseur logiciel	5To/an	FluxoMet
Données traitées	LCMS	Projets TraceFinder multiples extensions	Propriétaire/spécifique du fournisseur logiciel	6To/an	FluxoMet
Données converties	LCMS	.mzml	Format ouvert de données spectrales avec un minimum de métadonnées	1To/an	Format ouvert, standard du domaine FluxoMet
Données traitées	pipeline de traitement	.csv, tsv, txt...	données tabulées, normalisées et mises en forme	1Go/an	FluxoMet
RMN 1D	RMN 500 & RMN 800	fid, 1r, 1i	Format ouvert de données spectrales avec metadonnées	1To/an	FluxoMet
RMN 2D	RMN 500 & RMN 800	ser, 1rr, 1ri, 1ir, 1ii	*Format ouvert de données spectrales avec métadonnées*	2To/an	FluxoMet
RMN metadata	RMN 500 & RMN 800	txt	*Metadonnées, paramètres d'aquisition et de processing*	100 Mo/an	FluxoMet
Robot	Robot culture + robot prep	.log	Fichier text Fichier log	Go/an	FluxoMet
Robot	Robot culture	.txt	Fichier text Pegasus user data	100 Mo/an	FluxoMet
Robot	Robot culture	.csv	Fichier text data OD, pH, O2, pippetage	100 Mo/an	FluxoMet

Developpement logiciels

1a. Comment de nouvelles données seront-elles recueillies ou produites et/ou comment des données préexistantes seront-elles réutilisées ?

Les logiciels de la plateformes MetaToul sont issues des développements de la plateformes. Différents langages codées peuvent être utilisés comme Python, R.
Les logiciels existants font l'objet de maintenance et mises à jour régulières. Le logiciel git est utilisé pour réaliser le contrôle de version de ces logiciels.

1b. Quelles données (types, formats et volumes par ex.) seront collectées ou produites ?

Les codes sources (fichiers texte) des logiciels.

2. Documentation et qualité des données

Jeu de données MS, RMN, brutes (raw) et retraités

2a. Quelles métadonnées et quelle documentation (par exemple méthodologie de collecte et mode d'organisation des données) accompagneront les données ?

Plusieurs types des métadonnées seront traités :

Concernant les **données d'origines instrumentales**, le plan distingue deux grandes catégories de métadonnées :

Les métadonnées de projets de recherche. Elles correspondent principalement aux données administratives du projet scientifique, son plan d'expérience, les espèces biologiques ou les types d'échantillons. Elles sont collectées par les composants du système d'information de l'infrastructure MetaboHUB (portail MAMA) ainsi que par le système de management de la qualité ISO9001_NFX50-900 de **MetaToul** tel que la proposition d'études et de recherche, les cahiers de laboratoire ou la fiche de suivi de projet, les fiches échantillons, les revues de projet.

Les métadonnées de la production de la donnée. C'est l'ensemble des propriétés des méthodes d'acquisition (métadonnées de chromatographie, de l'analyseur de masse ou RMN, ...). Ces métadonnées sont stockées sur les machines de pilotage des instruments et dans le cas des standards chimiques peuvent être captées via une solution logicielle interne. L'ensemble des informations des acquisitions sont également enregistrées dans les cahiers de laboratoire à l'échelle de chaque membre de **Metatoul** ainsi que dans la fiche de suivi de chaque projet et la fiche échantillon.

Des métadonnées en lien avec les fichiers de données sont aussi capturées selon le standard Dublin core par les systèmes utilisés pour le stockage et la sauvegarde des données.

Des documents internes aux systèmes qualité des plateaux de **Metatoul** décrivent la collecte et l'organisation des données instrumentales et métadonnées scientifiques des projets (exemple : structuration des répertoires contenant les fichiers de données).

Pour les **données générées par le traitement de données**, les métadonnées résultantes vont varier selon le type d'outils utilisés pour le traitement. Ainsi, les membres de **Metatoul** distinguent les processus de traitement suivants :

- Des gestionnaires de flux de travail pour le traitement des données acquises sur les instruments. Ces environnements de traitement comme la plateforme en ligne Galaxy permettent de tracer et de reproduire les flux déjà établi et génèrent un ensemble de métadonnées relative aux outils utilisés, les paramètres fixés et les données d'entrées et de sorties produites.
- Des logiciels commerciaux, suivis dans la fiche suivi de chaque projet ou le cahier de laboratoire
- Des outils en cours de développement, et répondant aux besoins des projets analytiques de R&D et aux particularités de leurs données, suivis dans la fiche suivi de chaque projet ou le cahier de laboratoire.
- L'interprétation des données spectrales conduisant à la proposition d'une structure chimique

En lien avec la diversité des outils utilisés, les membres de **Metatoul** s'inscrit dans la dynamique des travaux en cours de l'infrastructure MetaboHUB sur « le rapport de résultats » et l'établissement d'un thésaurus commun.

2b. Quelles mesures de contrôle de la qualité des données seront mises en œuvre ?

Concernant les **données d'origines instrumentales**

Les membres du projet bénéficient de l'expérience de l'infrastructure MetaboHUB en la matière. Le workflow analytique comprend notamment une intégration de contrôles dans les séries (calibrant, QA, QC, blancs expérimentaux, standards internes) ainsi qu'un contrôle qualité des données générées (exemple : observation des pools en temps réel et pipeline de traitement à posteriori, monitoring des standards internes marqués), décrits des chaque mode opératoire appliqué.

Au niveau des **données générées par traitement de données**
Le contrôle repose sur l'expertise des agents en charge du traitement de données au vu de l'hétérogénéité des méthodes employés. Nous pouvons cependant distinguer
Pour les approches quantitatives (absolues) :

- Le R2 des droites de calibrations doit être supérieur ou égal à 0.99
- 6 points de gamme au minimum encadrant les mesures des échantillons à quantifier
- Des blancs expérimentaux (y compris un tube de standards internes) sont systématiquement analysés

Pour les approches quantitatives (relatives) :

- Des blancs expérimentaux (y compris un tube de standards internes) sont systématiquement analysés
- Les standards internes doivent être presents dans tous les échantillons

Pour les approches globales:

- Les QCs doivent être centré sur les analyses en composante principale

Developpement logiciels

2a. Quelles métadonnées et quelle documentation (par exemple méthodologie de collecte et mode d'organisation des données) accompagneront les données ?

Les logiciels sont accompagnés à minima d'un fichier README sur leur dépot afin d'expliciter le protocole d'installation et d'usage.
Une documentation en ligne est rédigée pour l'usage des logiciels demandant des compétences avancées (ligne de commande, analyses).

2b. Quelles mesures de contrôle de la qualité des données seront mises en œuvre ?

Les logiciels sont validés avec des jeux de données test.

3. Stockage et sauvegarde pendant le processus de recherche

Jeu de données MS, RMN, brutes (raw) et retraités

3a. Comment les données et les métadonnées seront-elles stockées et sauvegardées tout au long du processus de recherche ?

Les données expérimentales produites par **Metatoul** sont qualifiées de durables et non reproductibles. Les plateformes analytiques sécurisent toutes les données brutes produites en les stockant sur des serveurs informatiques adaptés, hébergés dans les salles de serveurs informatiques des instituts représentés dans le consortium. La sécurité des données est actuellement assurée par le niveau de sécurité des serveurs de l'INRAE, des universités, du CEA, de l'INSERM et de l'INSA qui les hébergent.

Plateau MetaToul-FluxoMet :

Les données d'acquisition (ou données brutes ou raw data) issues des spectromètres de masse haute résolution, des spectromètres RMN ou des robots de culture ou de préparation d'échantillons sont directement archivées/stockées sur les briques de stockage d'AgroDataRing (ADR) dès la fin de leur production dans un dossier dédié RAW_DATA

Pour effectuer le traitement des données brutes, celles-ci seront copiées en local sur les serveur T:/ ou U:/ du site hébergeur INSA dans un répertoire temporaire actif. Dans ce même répertoire se trouveront également toutes les métadonnées associées au projet (fiche échantillons, sample listing, modes opératoires, proposition d'études, suivi de projet etc..).

A la fin du processus de traitement et.ou à la clôture d'un projet, et après envoie des résultats au collaborateur ou demandeur, les données traitées et les métadonnées associées seront archivées sur ADR dans des répertoires dédiés (TREATED_DATA et METADATA).

Pour consulter des données archivées d'un projet en cours ou terminé, un import des données sera fait sur le serveur U:/, et à la fin de la consultation ou du retraitement, ces nouvelles données seront transférées sur ADR.

Enfin, concernant les données et métadonnées de projets archivés depuis deux ans ou plus, elles passeront en archivage froid sur bande magnétique.

Plateau MetaToul-Lipidomique

Les données d'acquisition (ou données brutes ou raw data) issues des instruments analytiques à détection de flamme (FID), spectromètres de masse haute et basse résolution ou des robots de préparation d'échantillons sont directement archivées/stockées sur les briques de stockage d'AgroDataRing (ADR) dès la fin de leur production dans un dossier dédié RAW_DATA

Pour effectuer le traitement des données brutes, celles-ci seront copiées en local sur les ordinateurs bureau des ingénieurs. Ils sont sauvegarde tous les jour à 12h30 sur le NAS lipidomique hébergé à l'I2MC Inserm 1297. Dans ce même répertoire se trouveront également toutes les métadonnées associées au projet (fiche échantillons, sample listing, modes opératoires, proposition d'études, suivi de projet etc..). Chaque données associé à un projet (données retraitées et métadata) est également présent dans le fichier du projet sur le Share point de Metatoul/Lipidomique/projets

Plateau MetaToul-AgroMix

Plateau MetaToul-Axiom

Les données brutes issues des spectromètres de masse, des spectromètres RMN ou du robot de préparation d'échantillons sont stockées sur les PC de pilotage des instruments au moins 1an, sauvegardées automatiquement et quotidiennement (données MS) sur le datacenter INRAE de Toulouse au moins 3ans. Les données du datacenter INRAE sont sauvegardées localement durant 60jours.

Les données traitées sont stockées sur les PC de chaque membre du plateau, lui-même synchronisé sur le datacenter INRAE de Toulouse pendant au moins 3ans. Les données du datacenter INRAE sont sauvegardées localement durant 60jours.

Les métadonnées du projet de recherche, sont stockées sur le PC du chef de projet, et sauvegardées comme les données traitées.

Les métadonnées d'acquisition des données enregistrées dans chaque donnée brute sont stockées et sauvegardées comme les données brutes, et les métadonnées enregistrées sur les documents papiers (cahiers de laboratoire, fiche suivi échantillon) sont conservées au moins 3 ans dans l'unité Toxalim.

Les métadonnées de traitement des données sont stockées sur le PC du membre de Metatoul ayant traiter les données, et sont sauvegardées comme les données traitées.

Plateau MetaToul - FluxoVivo

Les données d'acquisition (données brutes) issues du spectromètre de masse à haute résolution sont stockées/copiées manuellement sur le serveur du Restore dans un dossier dédié (public) accessible par l'ensemble des membres de Restore. Pour effectuer le traitement des données brutes, le traitement se fait sur les données copiées, soit depuis la station d'analyse soit depuis les postes de travail ayant les logiciels de traitement.

Dans un autre dossier spécifique au plateau FluxoVivo sur le serveur de Restore se trouveront les métadonnées associées au projet (fiche échantillons, sample listing, modes opératoires, etc..) ainsi que les données retraitées. Ce dossier est accessible uniquement par les membres de l'équipe.

Plateau MetaToul-MetExplore :

Les données sont hébergées sur ces serveurs locaux qui sont répliqués sur le dispositif de sauvegarde AgroDataRing (ADR). Cette solution repose sur plusieurs datacenter délocalisés qui permettent une sauvegarde jusqu'à 5 ans voir plus en fonction des besoins des projets.

La nomenclature des fichier de sauvegarde est normalisé au niveau de l'ensemble de la platforme:

Nomduplateau_N°MAMA_Analyse_RawData

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 **Metatoul** est associé.

En 2021, la production annuelle a atteint les 15 To/an. L'arrivée des instruments financés par le projet **METEX+** va considérablement augmenter les besoins en stockage de données au sein du consortium d'ici 2024 avec leur mise en service progressive. L'extension de la plateforme de stockage et de sauvegarde de MetaboHUB dont est partie prenante Metatoul est déjà en discussion en lien avec les DSI (Direction des Services Informatiques) de INRAE

3b. Comment la sécurité des données et la protection des données sensibles seront-elles assurées tout au long du processus de recherche ?

La plateforme **Metatoul** ne produit pas de données classées comme sensibles. Il est à noter que l'accès aux données est soumis à une authentification personnelle. Seuls les analystes de Metatoul ont accès à ces données. En cas de collaboration, un partage peut être mis en place, en interne via authentification ou en externe via un service de partage institutionnel (avec contrôle des accès).

Les données sauvegardées sont hébergées actuellement sur un site géographique différent de celui du stockage en fonction des différents sites.. En cas d'incidents, les données sauvegardées peuvent être récupérées via des outils internes. En cas de données plus anciennes et mises hors ligne, la récupération est assurée par l'administrateur à partir d'espaces secondaires.

Developpement logiciels

3a. Comment les données et les métadonnées seront-elles stockées et sauvegardées tout au long du processus de recherche ?

Les codes sources de logiciels sont sauvegardés en local et en ligne sur un dépot gitlab institutionnel (forgemia.inra.fr, gitlab).

3b. Comment la sécurité des données et la protection des données sensibles seront-elles assurées tout au long du processus de recherche ?

Les codes sources ne sont pas considérés comme données sensibles.

4. Exigences légales et éthiques, codes de conduite

4a. Si des données à caractère personnel sont traitées, comment le respect des dispositions de la législation sur les données à caractère personnel et sur la sécurité des données sera-t-il assuré ?

Les exigences légales et éthiques et le code de conduite sont alignés avec les informations contenues dans l'accord de consortium de l'infrastructure MetaboHUB (Avenant n°5 – 2022) **dont fait partie la plateforme Metatoul**, ainsi que par les accords de consortium pour chaque projet contractualisé avec des Tiers.

Les composants du système d'information scientifiques disposent des données d'utilisation informant les utilisateurs de ce qui est fait par exemple des informations de connexion recueillies. Deux exemples avec les composants PeakForest et MetExplore sont disponibles en ligne.

Le système d'information de gestion de projets MetaboHUB (MAMA) est le seul composant soumis au RGPD et inclus les fonctionnalités nécessaires.

Il est à souligner que les membres **de la plateforme Metatoul** ne gèrent pas de métadonnées personnelles ou sensibles. Si des métadonnées, par exemple de patients, sont nécessaires pour conduire des étapes d'acquisition ou de traitement de données, les processus d'anonymisation ou la pseudonymisation sont gérés en amont par les usagers et les porteurs de projets scientifiques demandeurs des analyses, et décrits dans les accords de consortium desdits projets.

Une réflexion sur le chiffrement des données est en cours au sein de l'infrastructure MetaboHUB.

Enfin, les membres **Metatoul** étant membres d'une infrastructure de recherche labellisée par INRAE (MetaboHUB), ils sont, de fait, adhérents de sa charte avec des engagements sur l'éthique et la déontologie des projets scientifiques menés.

4b. Comment les autres questions juridiques, comme la titularité ou les droits de propriété intellectuelle sur les données, seront-elles abordées ? Quelle est la législation applicable en la matière ?

Les aspects juridiques (titularité ou les droits de propriété intellectuelle sur les données) sont traités et validés pour chaque plateau de MetaToul à travers l'accord de l'unité d'adossement du plateau.

La propriété intellectuelle des données 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.

Projet de prestation :

Les données brutes et les données traitées appartiennent au partenaire avec lequel la plateforme Metatoul réalise le projet de prestation concerné. Les méthodes appliquées et l'ensemble des métadonnées associées au projet appartiennent à la plateforme Metatoul

Projet de collaboratif (ANR, BPI ...) :

La propriété des données brutes, des données traitées, et des métadonnées sera partagée selon le contrat rédigé entre les partenaires du projet et la plateforme Metatoul.

Projet de mise à dispo

L'ensemble des données et métadonnées appartiennent au partenaire avec lequel la plateforme Metatoul réalise le projet de mise à disposition

---

4c. Comment les éventuelles questions éthiques seront-elles prises en compte, les codes déontologiques respectés ?

La question des erreurs de traitement de données (exemple : erreur de calculs, algorithme défaillant dans un cas particulier, ...) est en cours de discussion au sein de la plateforme MetaToul et de l'infrastructure MetaboHUB.

Les questions éthiques relatives aux projets et échantillons analysés sont traités dans le cadre des accords de consortium avec les Tiers.

La plateforme n'est pas responsable de l'utilisation faite par les demandeurs des résultats des analyses. Toute utilisation partielle ou inappropriée ou toute interprétation dépassant les conclusions des rapports émis par la plateforme ne saurait engager la responsabilité de l'institut de tutelle de la plateforme Metatoul

---

5. Partage des données et conservation à long terme

Jeu de données MS, RMN, brutes (raw) et retraités

5a. Comment et quand les données seront-elles partagées ? Y-a-t-il des restrictions au partage des données ou des raisons de définir un embargo ?

La plateforme Metatoul souhaite s'inscrire dans la dynamique impulsée par le MESRI, les agences de financement, les établissements de recherche et les universités en termes d'Open data. Selon les textes en vigueur actuellement, deux conditions préliminaires sont actuellement nécessaires pour diffuser les données selon les principes de « Open Data » : des données réalisées dans le cadre de la mission de service public des établissements et des données achevées.

Les données produites dans le cadre des projets de R&D développés sur la plateforme ne sont pas concernées par le partage sauf dans le cas de développements d'outils de traitement de données ou de bases de données dans le consortium et nécessitant un jeu de données de démonstration ou de validation.

Dans ce contexte, nous distinguons le partage public par le dépôt des données sur un dépôt public de référence du partage par accord avec un nouveau partenaire désirant exploiter des données déjà acquises.

Concernant le partage public :

La plateforme Metatoul incite lors de 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. La rédaction d'un "data paper" en collaboration avec le porteur et sa publication dans un journal à comité de relecture international seront également proposés pour accélérer et valoriser le processus de partage des données. Si aucun accord n'est possible ou que le porteur ne répond pas, les données seront partagées selon les modalités décrites ci-dessus (sauf cas des données issues de l'étude de cohortes humaines en cours de réflexion – voir si dessous).

Concernant le partage par accord :

Si besoin, ce point sera défini, au cas par cas, dans les accords de consortium des projets scientifiques demandeurs des analyses. Le partage dans un nouveau projet ancillaire ne se fera pas sans avoir un accord de la part de l'ancien consortium (hors données publiques).

Relativement aux données spécifiques en lien avec les **cohortes humaines**, une action est en cours au sein de l'infrastructure MetaboHUB et en collaboration avec l'INSERM pour réfléchir au cadre à donner pour être en conformité avec les accords passés et futurs obtenus avec les porteurs de ces cohortes mais aussi se conformer à la législation liée au Règlement Général sur la Protection des Données (RGPD) et le cas des données personnelles.

Enfin, l**a plateforme** n'assure pas et ne propose pas de service d'archivage long terme de données (supérieur à 10 ans).

5b. Comment les données à conserver seront-elles sélectionnées et où seront-elles préservées sur le long terme (par ex. un entrepôt de données ou une archive) ?

Pour le moment, l**a plateforme Metatoul** s'engage à conserver l'ensemble des données produites pour une période de 5 ans. Cette période sera affinée dans les prochaines version du plan selon les domaines d'origine des données (humain ou plante).

5c. Quelles méthodes ou quels outils logiciels seront nécessaires pour accéder et utiliser les données ?

La plateforme MetaToul n'est pas concernée par le long terme (sauvegarde des données maximum 5 ans puis responabilité du demandeur/projet). Durant la période de stockage assurée par MetaToul (5 ans), la plateforme assure le maintient de l'ensemble des logiciels nécessaire à la lecture et l'exploitation de ces données

---

5d. Comment l'attribution d'un identifiant unique et pérenne (comme le DOI) sera-t-elle assurée pour chaque jeu de données ?

La plateforme Metatoul recommande l’attribution d’identifiants aux jeux de données de type DOI. Il n’est pas envisagé actuellement d’attribuer systématiquement de DOI à la production d’un jeu de données. Cette attribution reste à la charge du scientifique en charge du projet scientifique demandant des analyses.

Les ressources recommandées par MetaToul sont actuellement *Zenodo* (https://zenodo.org/), Data Gouv (https://www.data.gouv.fr) et *Data INRAE* (https://data.inrae.fr/).

Developpement logiciels

5a. Comment et quand les données seront-elles partagées ? Y-a-t-il des restrictions au partage des données ou des raisons de définir un embargo ?

Les logiciels en productions sont partagés à la communauté scientifique sous licence GPL? CC-BY? via les dépôt public (forgemia, gitlab, github).
Les logiciels en développement ne sont partagés qu'en interne.

5b. Comment les données à conserver seront-elles sélectionnées et où seront-elles préservées sur le long terme (par ex. un entrepôt de données ou une archive) ?

Pas de conservation sur long terme mise en place. Les logiciels librement accessibles sur dépots git sont archivés automatiquement par des sites externes (https://archive.softwareheritage.org)

5c. Quelles méthodes ou quels outils logiciels seront nécessaires pour accéder et utiliser les données ?

Chaque logiciel précise dans sa documentation leur protocole d'installation et d'usage.

5d. Comment l'attribution d'un identifiant unique et pérenne (comme le DOI) sera-t-elle assurée pour chaque jeu de données ?

HAL ? zenodo?

6. Responsabilités et ressources en matière de gestion des données

6a. Qui (par exemple rôle, position et institution de rattachement) sera responsable de la gestion des données (c'est-à-dire le gestionnaire des données) ?

La répartition des responsabilités pour la gestion des données s’effectue entre les porteurs des projets scientifiques et les responsables scientifiques des projets scientifiques, membres de la plateforme Metatoul.
Les décisions d’ouverture des données se prennent au niveau des établissements et ds financeurs.
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

plateformes.

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 membres **de Metatoul** également sous la responsabilité des responsables de plateformes.

Le stockage et sauvegarde, archivage et partage des données produites sont assurés par les administrateurs systèmes de la plateforme Metatoul

Chaque Plateau de la plateforme MetaToul possde un referent gestion de données qui sont (à janv 2023) :

- Metatoul-Lipido: Pauline Le-Fadouer-(INSERM)
- Metatoul-AgromiX:Guillaume Marti (UT3)
- Meatoul-FluxoMet: Floriant Bellvert (CNRS) et Etienne RIFA (INRAE)
- Metatoul-Axiom: Emilien Jamin (IR-INRAE)
- MetaToul -FluxoVivo Spiro Khoury (UT3) et Mathieu Vigneau (CNRS)
- MetaToul-MetExplore Fabien Jourdan (INRAE)

Guillaume Marti assurera le lien entre les sauvegarde local et distantes avec la solution ADR

Dans le cadre de la production de données pour des projets scientifiques menés en collaboration, la responsabilité de la gestion sera répartie entre les partenaires. Les membres **de Metatoul** s'engagent 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.

6b. Quelles seront les ressources (budget et temps alloués) dédiées à la gestion des données permettant de s'assurer que les données seront FAIR (Facile à trouver, Accessible, Interopérable, Réutilisable) ?

Concernant l'infrastructure informatique, **la plateforme Metatoul** s'appuie sur les investissements réalisés (budget + ressources humaines) dans le cadre du projet collaboratif AgroDataRing soutenu et impliquant l'infrastructure nationale MetaboHUB avec notamment :

Un budget annuel pour le renouvellement des serveurs de stockage/sauvegarde

Des ETPs en administration système

Des ETPs en développement informatique pour la maintenance des outils

Concernant l'ouverture des données, **la plateforme** s'appuie sur la politique d'investissement de l'infrastructure nationale MetaboHUB en termes de formation de ses personnels à la science ouverte et aux principes FAIR, et à l'utilisation des outils permettant cette ouverture des données.

Les processus d'ouverture des données et leur dépôt sur les entrepôts de référence ont des coûts en ressources humaines importants et sont non supportés par les membres **de la plateforme Metatoul.**