DMP du projet "A Tool for the ExploRation and Integration of omiCS data"

Plan de gestion de données créé à l'aide de DMP OPIDoR, basé sur le modèle "INRAE - Trame générique projet v1" 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 "A Tool for the ExploRation and Integration of omiCS data"
Version	Version intermédiaire
Langue	fra
Date de création	2021-01-22
Date de dernière modification	2021-08-30
Licence	Creative Commons Attribution Share Alike 4.0 International
Documents (publications, rapports, brevets, plan expérimental….), sites web associés	• Site web (outil ASTERICS) : https://asterics.miat.inrae.fr
Plans de gestion en lien avec le projet	• PGD Genotoul Bioinfo : 9242

Renseignements sur le projet

Titre du projet A Tool for the ExploRation and Integration of omiCS data

Acronyme ASTERICS

Résumé

Dans ce projet, nous proposons de combiner les compétences spécifiques des acteurs du projet (plateforme biostatistique de Genotoul, plateforme bioinformatique de Genotoul, Hyphen-Stat) pour développer des outils en ligne facilitant la préparation, l'analyse statistique et l'intégration des données issues des technologies à haut débit (données NGS) mais aussi l'accès aux données et aux chaînes de traitement ainsi que leur ré-utilisation en mettant l'accent sur l'**interopérabilité des données et des outils**. Des scénarii types potentiellement complexes enchaînant outils diagnostiques dont les sorties sont commentées de manière automatique avec mise en œuvre d'approches pour l'exploration de données ou la prédiction utilisant les choix issus des recommandations des outils diagnostiques seront intégrés à l'outil. Une validation sera réalisée sur deux cas d'études, un issu de données publiques et le second lié à une question d'intérêt en élevage sur la mortalité périnatale des animaux.

Sources de financement

- Région Occitanie :

Date de début 2020-09-15

Date de fin 2022-09-14

Partenaires

- National Research Institute for Agriculture, Food and the Environment ()
- Toulouse III University - Paul-Sabatier ()
- Hyphen-Stat ()

Produits de recherche :

1. Logiciel ASTERICS (Logiciel)

Contributeurs

Nom	Affiliation	Rôles
Nathalie Vialaneix		• Coordinateur du projet • Personne contact pour les données • Responsable du plan de gestion de données
Elise Maigne		

DMP du projet "A Tool for the ExploRation and Integration of omiCS data"

Informations sur le plan de gestion

Les responsables du PGD sont :

- Nathalie Vialaneix (PI du projet)
- Élise Maigné (participante)

Université de Toulouse, INRAE, UR MIAT, F-31320, Castanet-Tolosan, France

22/01/2021

version 0.2

version 0.1

Informations sur le projet

Projet Région Occitanie « Plateformes régionales de recherche et innovation »

Région Occitanie

Projet Région Occitanie « Plateformes régionales de recherche et innovation »

20008788

ASTERICS

A Tool for the ExploRation and Integration of omiCS data

INRAE, France

Université Paul Sabatier, Toulouse, France (partenaire)
Hyphen-stat https://www.hyphen-stat.com/ (partenaire)

Université de Toulouse, INRAE, UR MIAT, F-31320, Castanet-Tolosan, France
198717966P : MIAT Mathématiques et Informatique Appliquées Toulouse, INRAE, Toulouse, France

15/09/2020-14/09/2022 (2 ans)

Présentation générale des données du projet

Les données gérées et générées par le projet sont de deux types : 1/ les **données « cas d'études »** utilisées dans le cadre du projet pour tester et illustrer le développement du logiciel ; 2/ les **données utilisateurs** qui sont, d'une part les données personnelles utilisateurs requises pour la création du projet (email de l'utilisateur) et, d'autre part, les données chargées par les utilisateurs sur le logiciel lorsqu'ils réalisent leurs analyses ; 3/ le logiciel.

Données « Cas d'études »

1. **Données issues de l'entrepôt public TCGA** (https://portal.gdc.cancer.gov/). C'est la version disponible pour le cas d'étude décrit ici et complètement disponibles à ce lien sous forme de fichiers CSV et TXT avec les données d'origine et de quelques données complémentaires générées pour des besoins de test par le script R aussi fourni à ce lien. Les données ont été téléchargées en septembre 2020 et extraites au format CSV par Nathalie Vialaneix.
2. **Données issues du projet ANR « PORCINET »** (ANR-09-GENM005). Ces données ont été fournies par Laurence Liaubet (GenPhySE, INRAE) sous la forme de 5 fichiers CSV correspondant à des données communes à 50 individus (métadonnées, protéome du muscle, protéome du foie, transcriptome du muscle, transcriptome du sang) et de 4 fichiers correspondant à des données communes à 444 individus (métadonnées, métabolome liquide amniotique, plasma et urine). Ces données ont été acquises durant le développement fœtal de porcelets en fin de gestation.

Données utilisateurs

Les données utilisateurs ne sont pas utilisées pour le développement du projet. Elles restent la propriété exclusive de l'utilisateur. Les données personnelles seront collectées, stockées, gérées et effacées comme décrit dans les Privacy Policy du site web ASTERICS. Les données chargées par les utilisateurs seront stockées provisoirement sur le serveur du logiciel après son déploiement le temps de leur utilisation pour l'analyse. Ces données sont détruites un mois après la dernière visite du projet.

Le logiciel

Les données générées sous la forme d'un logiciel correspondent au code source du logiciel et à une image (docker ou singularity) à générer à partir de ce code source. Elles sont produites par le projet.

Droits de propriété intellectuelle

Les données et produits du projet comprennent :

- Le **cas d'études TCGA**. Les données de ce cas d'étude, disponibles publiquement, sont issues du projet The Cancer Genome Atlas (TCGA) et restent la propriété de ce projet. Les conditions d'utilisation de ces données sont celles décrites sur le portail et reprises dans la diffusion du jeu de données ASTERICS, à savoir :
  - les utilisateurs ne doivent pas essayer d'identifier les individus les participants des études dans lesquelles ces données ont été collectées ;
  - les utilisateurs doivent systématiquement citer le jeu de données précis dans toutes les communications orales ou écrites concernant ces données ;
  - ces données doivent être utilisées à des fins de recherche uniquement.
- Le **cas d'études PORCINET**. Les données de ce cas d'étude sont la propriété d'INRAE et restent la propriété d'INRAE. Une partie de ses données est disponible publiquement et une partie est à usage interne aux collaborateurs du projet PORCINET jusqu'à publication des résultats liés à ces données.
- Le **logiciel ASTERICS**. Le logiciel est la propriété d'INRAE et sera diffusé sous licence libre GPL3. Il pourra également faire l'objet d'une déclaration d'invention de type « déclaration d'invention œuvre informatique ».
- Les **données utilisateurs** restent l'intégrale propriété des utilisateurs et ne seront pas consultées, utilisées ni diffusées.

Le projet dépend de langages de programmation et de librairies (ou « packages ») diffusés sous licence libre (principalement GPL 2 ou supérieure). Ces dépendances ne créent pas de restriction d'utilisation ou de diffusion particulières mais nécessitent la publication de l'outil sous licence GPL.

Confidentialité

Aucune des **données « cas d'études »** de ce projet n'est confidentielle : les données du cas d'études TCGA sont publiques. Les données du cas d'études PORCINET ont vocation à devenir publiques après valorisation de celles-ci dans des publications en biologie (et une partie de ces données a déjà été diffusées dans des entrepôts publics). Les données du cas d'études TCGA seront incluses directement dans le logiciel final alors que les données du cas d'études PORCINET seront uniquement utilisées pour l'illustration de l'utilisation du logiciel via une documentation spécifique et ne seront pas diffusées via ce logiciel.

Les **données utilisateurs** sont confidentielles. Les données téléchargées par l'utilisateur d'ASTERICS sur le serveur restent son entière propriété, ne sont accessibles que par lui via le lien unique qui lui a été fourni et ont une durée de vie limitée à 1 mois sur le serveur. Les données personnelles confidentielles récoltées par ASTERICS sont décrites dans les Privacy Policy de l'outil.

L'utilisation de la version déployée en ligne du logiciel prévoira le consentement éclairé des participants qui auront à charger leurs données (ci-dessus désignées « données utilisateurs ») sur le serveur. Les détails de la sécurisation des données seront précisées dans la charte d'utilisation :

- Durée de vie : 1 mois d'inactivité pour les données personnelles et téléchargées d'un projet utilisateur.
- Sécurisation : l'accès aux données d'un projet utilisateur se fait par identifiant unique généré aléatoirement et fourni à l'utilisateur à la création de son projet. Les données téléchargées par l'utilisateur sur le projet ainsi que ses données personnelles sont stockées sur le serveur de production, dont l'accès est sécurisé et restreint aux administrateurs du projet (connexion par SSH).

Charte utilisateur d'ASTERICS.

Non concerné : les données utilisateurs ne seront jamais partagées.

Partage des données à l'issue du projet

Par contrat pour le financement de ce projet, le logiciel (code source) doit être partagé sous licence libre à l'issue du projet.

Données « Cas d'études »

1. **Données issues de l'entrepôt public TCGA** (https://portal.gdc.cancer.gov/). Ces données ont été partagées via le portail Data d'INRAE (DOI du portail : 10.14758/9T8G-WJ20 et DOI des données : 10.15454/YNMQUY). Elles seront aussi partagées à l'issue du projet par inclusion dans le logiciel comme données d'exemple et par utilisation dans le manuel utilisateur comme cas d'études exemple.
2. **Données issues du projet ANR « PORCINET »**. Ces données ne seront pas partagées par le projet à son issue. Elles seront toutefois utilisées pour la phase de test, le rapport interne final du projet et dans la documentation, comme illustration.

Données utilisateurs

Les données utilisateurs ne sont jamais partagées.

Logiciel

Les sources du logiciel seront partagées à l'issue du projet.

Les **données issues de l'entrepôt public TCGA** mises en forme par le projet peuvent être réutilisées pour :

- construire des formations pour l'utilisation du logiciel (probable),
- test du logiciel (probable),
- extraction d'informations nouvelles sur le cancer (peu probable car elles n'ont pas été extraites dans cette intention ici).

Toutefois, l'utilisation de ces données doit respecter les conditions d'utilisation de ces données, aussi rappelées dans la section « Droits de propriétés intellectuelles » de ce plan de gestion des données.

Les **sources du logiciel** pourront être utilisées pour déploiement du logiciel sur d'autres serveurs que le serveur de production du projet.

Les **données issues de l'entrepôt public TCGA** seront diffusées au format texte simple (CSV) ne nécessitant pas de recours à un logiciel spécifique.

Les **sources du logiciel** ne comprennent pas de code compilé et sont donc également lisibles sans logiciel spécifique.

Les **sources du logiciel** seront rendues disponibles sur la forge logicielle MIAT https://forgemia.inra.fr/asterics/asterics. L'image docker ou singularity du logiciel sera rendue disponible sur un dépôt public d'images comme celui de la forge MIA ou bien sur docker hub ou singularity hub.
Les **données issues de l'entrepôt public TCGA** ont déjà été rendues disponibles via le portail Data d'INRAE (DOI: 10.15454/YNMQUY). Elles seront aussi rendues disponibles dans les sources du logiciel et également, de manière indirecte, dans le déploiement du logiciel (par chargement du cas test sur l'interface à la demande de l'utilisateur).

Les **sources du logiciel** seront diffusées sous licence GPL3.
Les **données issues de l'entrepôt public TCGA** sont diffusées avec les restrictions de conditions d'utilisation imposées par le projet et précisées dans la partie « Droits de propriété intellectuelle » de ce plan de gestion des données et dans la partie Conditions du dépôt Data d'INRAE (DOI : 10.15454/YNMQUY). Ces conditions seront rappelées dans la charte utilisateur.

Les **sources du logiciel** seront rendues disponibles lors du déploiement du logiciel sur le serveur de production (à la fin du projet).
Les d**onnées issues de l'entrepôt public TCGA** ont été rendues disponibles à un an de projet (Août 2021) sur le portail Data d'INRAE (DOI : 10.15454/YNMQUY). Elles seront également disponibles lors du déploiement du logiciel sur le serveur de déploiement (prévu automne 2021) et dans la diffusion des sources du logiciel (à la fin du projet).

Les sources du logiciel et les données issues de l'entrepôt public TCGA seront partagées sans date de fin prévue.

Les sources du logiciel ne seront pas identifiées par un identifiant pérenne.
Les données du cas d'études TCGA sont diffusées via le lien pérenne de l'entrepôt Data d'INRAE (DOI : 10.15454/YNMQUY)

La demande de DOI pour le dépôt des données TCGA sur le Dataverse INRAE a été prise en charge par la responsable du projet, Nathalie Vialaneix.

Description et organisation des données

Concernant les données des cas d'études :

- pour le **cas d'études TCGA**, les données ont été acquises via l'utilisation d'une mise à disposition tierce de données pré-traitées (cas d'études du projet mixOmics / DIABLO) et ont été exportées aux formats CSV ou TXT. Des données complémentaires ont été générées par l'utilisation de scripts R pour les besoins du projet. Les détails de ces acquisitions sont disponibles sur le dépôt des données sur le portail Data d'INRAE (DOI : 10.15454/YNMQUY) ;
- pour le **cas d'études PORCINET**, l'acquisition des données est antérieure au projet ASTERICS qui n'est donc pas concerné par cette acquisition.

Concernant les **données utilisateur**, les données personnelles (email) seront fournies par l'utilisateur en remplissant un formulaire qui lui donne accès à l'outil (inscription). Les données téléchargées par l'utilisateur seront importées sur le serveur de production via un formulaire web sous forme de fichiers texte (formats TXT, CSV, TSV, ...). Elles serviront de base à une analyse menée par l'utilisateur et utilisant ASTERICS. Les données et les analyses seront conservées, sur le serveur, en données au format rda (données R). Elles seront automatiquement supprimées après 1 mois d'inactivité de l'utilisateur.
Concernant les **sources du logiciel**, elles seront mises à disposition sous la forme d'un conteneur (docker ou singularity).

Concernant les données des cas d'études :

- pour le **cas d'études TCGA**, les données sont documentées sur le dépôt des données sur le portail Data d'INRAE (DOI : 10.15454/YNMQUY) ;
- pour le **cas d'études PORCINET**, les données ne seront pas rendues publiques via le projet donc ne seront pas documentées. Elles seront toutefois décrites pour les besoins de l'illustration de ce cas d'études dans la documentation du logiciel.

Concernant les **données utilisateur**, elles ne sont pas concernées par la documentation.
Concernant les **sources du logiciel**, elles seront documentées par la diffusion d'une documentation développeur et d'une documentation utilisateur.

Concernant les données des cas d'études :

- pour le **cas d'études TCGA**, les métadonnées produites se conforment au schéma du Dataverse INRAE (auteur, références, type de données, ...) ;
- pour le **cas d'études PORCINET**, les données sont antérieures au projet et les données ne sont pas diffusé par le projet donc aucune métadonnée n'est produite.

Concernant les **données utilisateurs**, elle ne sont pas concernées par la diffusion.
Concernant les **sources du logiciel**, les métadonnées produites sont les auteurs, les versions du logiciel avec leurs dates de diffusion et les dépendances logiciels.

Les métadonnées du **cas d'études TCGA** ont été produites selon le schéma du portail Data d'INRAE. Les autres données ne sont pas concernées par la production de métadonnées.

Concernant les données des cas d'études :

- pour le **cas d'études TCGA**, les données ont été extraites sans modification et des fichiers complémentaires ont été produits par des scripts R qui ont été rendus publiques ;
- pour le **cas d'études PORCINET**, les données n'ont pas été modifées.

Concernant les **données utilisateur**, la structure du logiciel doit assurer la traçabilité et la reproductibilité des données issues des données chargées par l'utilisateur. L'utilisateur peut supprimer des analyses et données en cours de projet mais les résultats de toutes les analyses sont associées aux données ayant permis de les produire ainsi qu'aux options (arguments) des fonctions d'analyse qui les ont générées. Les différentes versions des analyses sont conservées de cette manière, jusqu'à leur suppression par l'utilisateur ou à expiration du projet utilisateur correspondant. Par ailleurs, l'utilisateur dispose d'un droit de consultation, rectification et suppression de ses données personnelles.
Concernant les **sources du logiciel**, la gestion des versions des sources est gérée par un dépôt Git hébergé sur la forge MIA. Celui-ci contient :

- une branche *master*, utilisée pour les sources envoyées en production ;
- une branche *dev*, utilisée pour les sources envoyées sur le serveur de développement, qui a vocation à être fusionnée dans *master* ;
- des branches correspondant à des fonctionnalités en cours de développement qui partent de *dev* et ont vocation à être fusionnées sur cette branche. Les demandes de fusion sont gérées par le développeur principal de la fonctionnalité correspondant à la branche et validées par un tiers expert dans cette fonctionnalité au sein du projet.

La qualité des sources logiciels est assurée par la mise en place de tests unitaires.

Stockage et sécurité des données

Concernant les données des cas d'études :

- pour le **cas d'études TCGA**, les données sont stockées sur le portail Data d'INRAE (DOI : 10.15454/YNMQUY) ainsi que sur l'espace projet du cluster Genotoul BIOINFO et sur le serveur de production ;
- pour le **cas d'études PORCINET**, les données sont provisoirement stockées sur l'espace projet du cluster Genotoul BIOINFO pour la durée du projet.

Concernant les **données utilisateur**, celles-ci seront stockées sur le serveur de production.
Concernant les **sources du logiciel**, elles sont stockées sur la forge MIA.

Le projet n'est pas concerné par cette question.

Concernant les données des cas d'études :

- pour le **cas d'études TCGA**, les données ont une volumétrie de l'ordre de 40Mo ;
- pour le **cas d'études PORCINET**, les données ont une volumétrie de l'ordre de 60Mo.

Concernant les **données utilisateur**, la volumétrie, pour chaque utilisateur, devrait être de l'ordre de grandeur de la volumétrie des

deux cas d'études. La volumétrie totale dépendra du nombre d'utilisateurs.
Concernant les **sources du logiciel**, la volumétrie du dépôt git de versionnement du code est de l'ordre de 20 Mo à un an de projet. La volumétrie finale ne devrait pas dépasser le double de cette valeur. La volumétrie d'une image docker ou singularity du logiciel devrait être de l'ordre de 1 à 2 Go.

Même réponse que pour la question sur le support de stockage des données.

Concernant les divers modes de stockages :

- le portail Data INRAE est hébergé par INRAE, 147 rue de l'Université, 75338 Paris cedex 07
- le cluster GenoToul BIOINFO et la forge MIA sont localisés dans le Data-Center « L'ARCHE DE DONNÉES Francis Sevila » localisé sur le centre INRAE Occitanie-Toulouse

---

Les règles de sécurité régissant les différentes entités hébergeantes sont régies par :

- la charte des infrastructures INRAE
- la charte de la plateforme GENOTOUL Bioinfo
- la politique de confidentialité de la plateforme GENOTOUL Bioinfo
- le plan de gestion des données (DMP) de la plateforme GENOTOUL Bioinfo (en cours de rédaction)

En ce qui concerne le Datacenter d'INRAE, un dispositif de contrôle d'accès (badges nominatifs) est mis en place, avec un accompagnement systématique pour les prestataires, ainsi que des caméras de surveillance 24h/24, dont une infrarouge. Chaque baie informatique est verrouillée par clé.
Au niveau immatériel, le Datacenter dispose de solutions de filtrage des flux en entrée et en sortie permettant de réduire les risques d'exposition des données et applique la politique de sécurité des systèmes d'information d'INRAE.

Les données ne font pas l'objet d'échange avec des tiers acteurs extérieurs au projet. Au sein du projet, l'entreprise Hyphen-Stat a accès au code source et aux données publiques du cas d'étude TCGA mais pas aux données utilisateur, ni aux données du cas d'étude PORCINET.

Une convention de partenariat sera établie entre les partenaires.

Les méthodes d'authentification pour l'accès aux données sont les méthodes d'authentification des services hébergeant (forge MIA, cluster GenoToul BIOINFO), par mot de passe ou connexion SSO, à l'exception du portail Data d'INRAE qui ne nécessite pas d'authentification pour la consultation et le téléchargement des données.

Ce projet n'est pas concerné par la production de données.
Concernant la politique de traçabilité du code et des données et analyses des utilisateurs :

- Nous utilisons un système de contrôle de version (Forge MIA, serveur git avec interface GitLab) pour le développement de l'application pour assurer sa traçabilité.
- Nous avons mis en place des procédure de tests des fonctions (en utilisant le package **testthat** de **R** pour les fonctions d'analyse en **R**) afin de s'assurer de la cohérence des résultats au fur et à mesure des évolutions de l'application.
- Les analyses des utilisateurs sont tracées par affichage d'un graphe dirigé acyclique (DAG) sur l'interface pour assurer leur reproductibilité dans l'interface. Un guide utilisateurs sera également publié pour conseiller l'utilisateur sur les analyses et options à utiliser selon de ses données et résultats.

Archivage et conservation des données après la fin du projet

Toutes les données diffusées dans le cadre de ce projet sont à conserver sur le long terme, à l'exception des données utilisateurs qui sont détruites après un mois d'inactivité du projet de l'utilisateur.

Concernant le **cas d'études TCGA**, les données sont archivées de manière pérenne sur le portail Data d'INRAE (DOI : 10.15454/YNMQUY) ;
Concernant les **sources du logiciel** et **l'image produite**, aucune politique d'archivage n'a été définie. Les sources n'ont pas vocation à être archivées mais à être versionnées.

Aucune durée maximale de conservation des données diffusées dans le cadre de ce projet n'est fixée, à l'exception des données utilisateurs qui sont détruites après un mois d'inactivité du projet de l'utilisateur.

Nathalie Vialaneix

Le volume des données est identique à celui décrit dans la partie « Quelle est la volumétrie prévisionnelle ? » de la section « Stockage et sécurité des données ».

Les coûts d'archivage sur le portail Data d'INRAE sont supportés par INRAE.