DMP du projet "Un portail opérationnel pour la production de produits biosourcés"

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 "Un portail opérationnel pour la production de produits biosourcés"
Version	Version initiale
Objet/périmètre du plan	Ce PGD décrit les règles qui vont être appliquées dans le projet "PEPR B-BEST Galaxy-BioProd" pour • publier et conserver les protocoles et données créés dans le cadre du projet (délivrable D1.1) • partager et conserver le code informatique produit et mettre à disposition les outils et workflow développés dans le cadre du projet (délivrables D2.1, D3.2, D4.1, D4.2, D4.3)
Domaines de recherche (selon classification de l'OCDE)	Industrial biotechnology, Computer and information sciences
Langue	fra
Date de création	2023-07-07
Date de dernière modification	2023-10-13
Licence	**Nom** Creative Commons Attribution Non Commercial 4.0 International **URL** http://spdx.org/licenses/CC-BY-NC-4.0.json

Renseignements sur le projet

Titre du projet	Un portail opérationnel pour la production de produits biosourcés
Acronyme	PEPR B-BEST Galaxy-BioProd
Résumé	L'objectif de ce projet ciblé est de fournir des outils et des ressources génériques et partagés pour concevoir, exécuter et suivre les projets des axes 1-3 du PEPR B-BEST. À cette fin, un portail centralisé fournira des outils logiciels ainsi que des ressources de calcul et de stockage. Au-delà des besoins du PEPR B-BEST, le système d'exploitation et de suivi proposé minimisera les temps et les coûts de

développement en R&D. En particulier, nous chercherons à intégrer données et plateformes robotiques pour accélérer le développement des (bio)catalyseurs, de la bioingénierie de souches et des (bio)procédés associés. Un tel système n'existe pas encore et, au-delà du projet actuel, devrait également impacter les communautés scientifiques de la biologie de synthèse, de la biocatalyse et des biotechnologies industrielles.

Afin de standardiser le processus long et coûteux (basé sur un cycle essais et erreurs) traditionnellement appliqué bio-ingénierie, la communauté de la biologie synthétique a développé et mis en œuvre au cours de la dernière décennie la technologie DBTL (Design-Build-Test and thereafter Learn). Aujourd'hui, le DBTL est le *modus vivendi* de toute biofonderie, notamment celles regroupées au sein de l'alliance mondiale des biofonderies. Cependant, il subsiste un manque de pratiques partagées et chaque biofonderie a déployé son propre pipeline. Pour pallier l'absence de pratiques communes, les outils et ressources informatiques développés et déployés dans le cadre du projet Galaxy-BioProd seront standardisés selon les principes FAIR (Findability, Accessibility, Interoperability and Reusability). Pour faciliter la FAIRification des données et des logiciels et offrir une solution facile à utiliser par tous, nous proposons d'utiliser le système de gestion de workflows Galaxy. Les gestionnaires de workflows scientifiques tels que Galaxy fournissent une plateforme ouverte pour effectuer des analyses de données liées à des protocoles expérimentaux pour tous les scientifiques, quelle que soit leur expertise informatique, ainsi que des calculs interopérables et reproductibles, quelle que soit la plateforme utilisée. Le système Galaxy est disponible via un navigateur internet, est largement utilisé (+8k outils), offre des ressources de formation en ligne et fournit une interface internet simple augmentant l'efficacité de ceux qui l'utilisent.

Nos développements au sein du gestionnaire de workflow Galaxy seront basés sur un ensemble de ressources déjà présentes dans le Galaxy ToolShed et couvrant l’analyse -omiques (y compris métabolomique), et l'ingénierie de souches en biologie synthétique et en ingénierie métabolique. Nous connecterons les outils existants aux bases de données omiques pertinentes et standardiserons leurs entrées et sorties de manière à ce qu'ils puissent être enchaînés pour former des workflows couvrant toutes les étapes du cycle DBTL. Pour répondre aux besoins des autres axes de ce PEPR, de nouveaux outils allant de TRL1 à TRL4 seront développés et déployés avec Galaxy, notamment (liste non exhaustive, du fondamental à l'appliqué) : la rétrosynthèse chimique/biochimique et à souches multiples, l'optimisation des séquences enzymatiques pour une stabilité évolutive accrue, l'apprentissage automatique actif pour la conception expérimentale automatisée, la modélisation multi-échelle des bioréacteurs industriels, les modèles hybrides pour le contrôle en ligne et la modélisation ACV des produits biosourcés.

Tous les outils adaptés et développés dans le cadre de ce projet comprendront une formation en ligne accessible depuis Galaxy et d'autres plateformes comme celles de l'infrastructure européenne Elixir.

Sources de

financement	• Agence Nationale de la Recherche : ANR-22-PEBB-0008
Date de début	2023-06-01
Date de fin	2027-05-31
Partenaires	• Institut national de recherche pour l'agriculture, l'alimentation et l'environnement https://ror.org/003vg9w96 • Centre de recherche CEA Paris-Saclay https://ror.org/01fv25t22 • CentraleSupélec https://ror.org/019tcpt25 • IFP Énergies nouvelles https://ror.org/03gcbhc33 • Institut national des sciences appliquées de Toulouse https://ror.org/01h8pf755 • Centre national de la recherche scientifique https://ror.org/02feahw73

Produits de recherche :

1. Stockage des protocoles et données suivant les règles et les principes FAIR (Jeu de données)
2. Outils bioinformatiques (Logiciel)

Contributeurs

Nom	Affiliation	Rôles
Cottret Ludovic	INRAE	• Personne contact pour les données (Outils bioinfo) • Responsable du plan
Faulon Jean-Loup - 0000-0003-4274-2953	INRAE - 201119643H	• Coordinateur de projet
Paes Gabriel	INRAE	• Personne contact pour les données (Protocoles)

DMP du projet "Un portail opérationnel pour la production de produits biosourcés"

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

Stockage des protocoles et données suivant les règles et les principes FAIR

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

Des protocoles et des données provenant de la littérature seront utilisés dans ce projet. La plupart seront libres d'utilisation mais certaines bases de données utilisées peuvent être propriétaires. Dans ce cas, cette restriction sera clairement indiquée dans un fichier de metadata décrivant les données, les protocoles et les logiciels les utilisant.

La standardisation des données et des protocoles utilisés fait partie du projet en soi donc elle se construira au cours du projet. Cependant, nous pouvons déjà dire que chaque donnée et protocole sera accompagné d'un fichier de metadata dont le format suivra une ontologie précise (ex: modèle ISA).

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

Les données seront de nature numérique et pourront être disposées dans des tableaux (format csv). Il s'agira de données de caractérisation :

- physique : taille de particules
- chimique : rendement d'hydrolyse, composition en monosaccharides, analyse élémentaire, codes InChI et SMILES
- bilans matière et énergie : débits d'entrée et de sortie, rendements
- inventaire de cycle de vie : quantités nécessaires pour l'unité fonctionnelle

Par leur nature, ces fichiers représenteront une volumétrie très faible (quelques kilo-octets).

Outils bioinformatiques

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

Différents langages informatiques seront utilisés au sein du projet comme Python, R, Java.
Les logiciels existants, utilisés au cours du projet, sont déjà l'objet de maintenances et mises à jour régulières. Le logiciel git est utilisé pour réaliser le contrôle de version de ces logiciels.

Nous nous appuierons, si nécessaire, sur des logiciels existants qui seront choisis en fonction de leur intérêt scientifique, de la compatibilité de leur licence avec les exigences du projet (licences de type Open Source sans restriction commerciale) et de leur maintenance (code maintenu par une communauté, mises à jour régulières, suivi des bugs...)

Galaxy, qui encapuslera les workflows et composants logiciels est un logiciel Open Source (license Academic Free License version 3.0.) dont le pilotage et la gouvernance sont documentés (https://galaxyproject.org/community/governance/).

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

La première production est le code source des logiciels. Le code source est l'ensemble des fichiers texte codant pour le fonctionnement du logiciel. Son volume représente le plus souvent moins d'1Go.
Les sources des logiciels peuvent être organisées en librairie avec des structures spécifiques selon le langage utilisé. Dans ce cas, elles peuvent alors être déposées sur des sites publiques (ex : maven, npm, Pypi) afin de faciliter leur diffusion au sein de la communauté.

Pour assurer la simplicité et la pérennité de l'utilisation du logiciel, ceux-ci seront fournis sous formes de conteneurs (ex : Conda, Docker, Singularity) qui contiennent tout l'environnement (système d'exploitation, librairies) nécessaire
au bon fonctionnement du logiciel. Ceci permet une installation simplifiée sur les ordinateurs individuels des utilisateurs
mais aussi sur les clusters de calcul qui l'utilisent. Dans ce cas, chaque nouvelle version du logiciel est placé dans un nouveau conteneur. Le volume de l'ensemble des conteneurs pour un logiciel peut ainsi facilement atteindre une dizaine de giga-octets. Ces conteneurs seront déposés sur la forge institutionnelle ou/et des plate-formes publiques pour faciliter leur
utilisation par l'ensemble de la communauté.

Les outils utiliseront les données en entrées et produiront les données en sortie ci-dessous :

- Retrosynthèse
  - entrée :
    - des structures moléculaires sous forme d'InChI (International Chemical Identifier)
    - des modèles métaboliques de type Genome-Scale Metabolic Models (GEMs) (SBML annoté)
    - une liste de plasmides et vecteurs (SBOL)
  - sortie : liste de voies métaboliques + enzymes catalysant les réactions (SBML annoté)
- Analyse des voies
  - entrée : liste de voies métaboliques (SBML)
  - sortie : liste de voies métaboliques (SBML annoté)
- Sélection des enzymes
  - entrée : séquences enzymatiques (FASTA)
  - sortie : séquences enzymatiques classées (FASTA)
- Ingénierie de souches
  - entrée : modèle métabolique (SBML)
  - sortie : modèle métabolique (SBML)

Également, des workflows seront construits en chainant les outils créés dans le projet. Ceux-ci seront sauvegardées dans le serveur galaxy financé par le projet.

Enfin, des tutoriels Galaxy Training seront produits comprenant du contenu au format texte, vidéo, jeux de données exemples et workflow.

2. Documentation et qualité des données

Stockage des protocoles et données suivant les règles et les principes FAIR

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

Chaque dossier partagé est accompagné d'un fichier texte du même nom avec le suffixe "README" qui contient les métadonnées selon le
format Dublin Core : Title, Creator, Subject, Description, Publisher, Contributor, Date, Type, Format, Identifier, Source, Language,
Relation, Coverage, Rights. A ces métadonnées de base pourront s'ajouter des métadonnées propres au type de protocole ou de donnée. La liste de ces métadonnées sera établie au cours du projet.

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

Tout fichier déposé sur l'espace partagé fait l'objet d'une double vérification : par la personne produisant le fichier et par le coordinateur qui vérifie le dépôt.

Les responsables du PGD vérifient également que chaque fichier est bien accompagné de son protocole d'obtention et

des fichiers README associés.

Pour les analyses chimiques, les limites de quantification sont calculées et chaque série d'analyses comprend des échantillons de références (standards).

Pour les inventaires de cycle de vie, une description de l'unité fonctionnelle et les frontières seront recueillis avec la provenance de données (base de données externe, calcul, mesure).

Outils bioinformatiques

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 dépôt afin de détailler 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).

Les données d'entrée seront téléchargées depuis les bases de données publiques. Aucune métadonnée ne sera liée aux données d'entrée.

Les workflows seront accompagnés de tutoriels au format "Galaxy Training Network", mis à disposition de la communauté sur un site dédié.

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

Les logiciels sont validés avec des jeux de données tests. Des tests unitaires peuvent être intégrés afin de garantir l'intégrité du logiciel au fil des développements. Des mesures de couverture de code par ces tests unitaires sont un gage de qualité du code fourni. Le contrôle de version est réalisé grâce au logiciel git. Il est demandé à ce que les codes sources soient bien commentés, et écrits en suivant au mieux les bonnes pratiques définies pour les langages de programmation utilisés. La plupart des langages et des éditeurs de code (ex : vscode, IntelliJ) mettent aujourd'hui à disposition des outils qui permettent d'assister le développeur pour respecter ces bonnes pratiques. Dans certains cas, ils proposent également des mesures de qualité du code.
Dans le cas particuliers des portails Web, plusieurs outils existent pour s'assurer de leur qualité. Des analyses en ligne sont ainsi réalisées pour mesurer les performances, l'accessibilité mais aussi la sécurité du site.

Les workflows seront également accompagnés de données de test. Toute nouvelle version du workflow devra passer ces tests pour être publiée sur le toolshed (règle IUC).

3. Stockage et sauvegarde pendant le processus de recherche

Stockage des protocoles et données suivant les règles et les principes FAIR

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

Un espace d'échange hébergé par l'INRAE (NextCloud) a déjà été mis en place pour rendre les données accessibles à tous les membres du projet.

Il est envisagé que les fichiers texte ou tabulés puissent être versionnés de la même façon que le code avec l'outils Git et soient déposés sur une forge logiciel. Ceci permettrait un partage des données et un suivi de leurs modifications. En utilisant les outils d'intégration continue, il sera ainsi possible de vérifier leur adéquation avec les principes FAIR.

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

L'espace d'échange NextCloud mis en place pour le projet a un accès authentifié. La forge logicielle (certainement celle de l'INRAE), potentiellement utilisée pour gérer les versions des protocoles et des données, sera également protégée par un accès authentifié. Pour ces deux ressources, il est possible de créer des groupes d'utilisateurs avec différents accès en fonction des données échangées.

Outils bioinformatiques

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 sur des machines locales et en ligne sur des dépôts institutionnels eux mêmes dupliqués sur les solutions publiques (forgemia.inra.fr, gitlab, github).

Les données générées seront stockées sur les serveurs hébergeant le serveur Galaxy au sein de l'IFB.

Les workflows Galaxy seront stockés sur le serveur qui hébergera la plate-forme Galaxy puis publiés sur le ToolShed (https://toolshed.g2.bx.psu.edu).

Les tutoriels Galaxy Training seront hébergés sur https://training.galaxyproject.org.

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, puisqu'ils seront sous une licence Open Source. Par contre, leur utilisation, en particulier à travers le Web, requiert souvent des précautions supplémentaires afin que les données traitées par le logiciel restents confidentielles. Des outils de diagnostique de code seront utilisés afin d'identifier les potentielles failles de sécurité.

La plate-forme Galaxy est protégée par un accès authentifié. L'infrastructure de l'IFB répond à l'état de l'art en termes de sécurité des données.

4. Exigences légales et éthiques, codes de conduite

Stockage des protocoles et données suivant les règles et les principes FAIR

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 ?

Pas de données personnelles traitées.

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 ?

L'accord de consortium établi entre les partenaires définit ces règles. Chaque partenaire reste propriétaire de ses données, en cas de question, elles seront remontées à l'équipe de gestion de projet qui évaluera la situation et échangera avec les référents opérationnels concernés des établissements impliqués.

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

Pas d'aspect éthique à prendre en compte.

Outils bioinformatiques

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 ?

Aucune donnée à caractère personnel ne sera traitée lors de ce projet.

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 ?

Le code produit au cours du projet sera mis sous licence Open Source (Apache 2.0, GPL, CECILL, ou CC-BY) afin d'en garantir les droits de propriété intellectuelle et leur libre réutilisation par d'autres projets de recherche.

Les restrictions d'utilisation dues aux licences non Open Source embarqués dans les librairies externes utilisées dans le cadre du projet seront précisés dans la documentation des logiciels.

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

Cette question est non applicable ici.

5. Partage des données et conservation à long terme

Stockage des protocoles et données suivant les règles et les principes FAIR

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 ?

Le versement sur les bases de données publiques (ex : data.gouv.fr) se fera en cas de publication, valorisation et exploitation des données.

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) ?

Les données et protocoles utilisés le long du projet, présents dans l'espace d'échange NextCloud, seront préservés le temps du projet. Ceux qui feront l'objet d'une publication, d'une valorisation ou d'une exploitation dans des outils de recherche seront préservés à long terme dans un dépôt de données public tel que data.gouv.fr.

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

Les données de caractérisations physiques et chimiques seront dans des tableaux de type CSV accessibles librement.

Les résultats des modèles seront dans des tableaux de type CSV. La reproduction des résultats nécessitera l'utilisation du code qui sera récupérable via une forge logicielle.

Les données des analyses de cycle de vie (ACV) seront également des tableaux de type CSV. Les inventaires peuvent être utilisés dans un logiciel ACV qui nécessitera d'autres bases de données payantes.

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

Un DOI sera attribué automatiquement aux jeux de données lors du dépôt sur data.gouv.fr.

Outils bioinformatiques

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 ?

Le projet est centré sur le développement et l'utilisation d'outils Open Source. Le code généré sera donc ouvert à la communauté dès qu'il atteindra une version utilisable et documentée.

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) ?

Les logiciels librement accessibles sur dépots git sont archivés automatiquement par des sites externes (https://archive.softwareheritage.org). Pour certains logiciels la création d’image Conda, Docker et/ou Singularity permet de fixer les différentes versions/releases . Ces images seront stockées sur la forge institutionnelle de l'INRAE et sur les dépôts publics.

Les logiciels embarqués dans Galaxy seront publiés sur le toolshed de Galaxy.

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. Il est différent selon le langage utilisé.

Les briques Galaxy seront accessibles à travers le portail Galaxy mis en place par l'IFB (https://usegalaxy.fr/).

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

SoftwareHeritage permet l'assignation d'un DOI (SHID) qui pourra être utilisé lors de la publication afin d'identifier de façon pérenne une version du logiciel.

6. Responsabilités et ressources en matière de gestion des données

Stockage des protocoles et données suivant les règles et les principes FAIR

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) ?

Chaque membre du projet est responsable de la maintenance de son protocole ou de ses données, au moins pendant le temps du projet.

Le responsable du PGD, Ludovic COTTRET, s'attachera au cours du projet à ce que la gestion des données suive les principes FAIR et à ce que le plan de gestion des données soit révisé régulièrement en fonction de l'avancement du projet.

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) ?

Une grande partie du projet est dédié à rendre FAIR les données et les outils. Une partie conséquente du budget et du temps sera donc allouée à cette tâche, par la nature même du projet.

De plus, le budget du projet intègre de l'équipement informatique et des ressources contractuelles dédiées spécifiquement à la gestion des données produites au cours du projet.

Outils bioinformatiques

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) ?

Chaque développeur est responsable de la maintenance de son code, au moins pendant le temps du projet.

L'objectif du projet est également de produire du code Open Source, pour le partager au sein de la communauté et que celle-ci puisse participer à son développement.

Le responsable du PGD, Ludovic COTTRET, s'attachera au cours du projet à ce que le code et les outils produits suivent les principes FAIR et à ce que le plan de gestion des données soit révisé régulièrement en fonction de l'avancement du projet.

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) ?

Une grande partie du projet est dédié à rendre FAIR les données et les outils. Une partie conséquente du budget et du temps sera donc allouée à cette tâche, par la nature même du projet.

De plus, le budget du projet intègre de l'équipement informatique et des ressources contractuelles dédiées spécifiquement à la gestion des données produites au cours du projet.