PGD DE LA PLATEFORME DE BIOINFORMATIQUE DES PLANTES DE L'URGI, PLANTBIOINFOPF

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

Table des matières

RENSEIGNEMENTS SUR LE PLAN

PLAN
Titre : PGD de la plateforme de bioinformatique des plantes, PlantBioinfoPF
Langue : Français
Date de création : 16/02/2021
Date de dernière modification : 04/06/2024

IDENTIFIANT DU PLAN
https://doi.org/10.15454/9HM5UI

LICENSE DU PLAN

RESUME
L'URGI, une unité de recherche INRAE, héberge une plateforme de bioinformatique (Plant Bioinformatics Facility - doi:10.15454/1.5572414581735654E12) qui soutient les activités de recherche en génétique et génomique sur les plantes. Les services de la plateforme couvrent la conception de bases de données, l'ingénierie logicielle, l'hébergement de logiciels, l'intégration de données et la formation. Les activités de la plateforme bénéficient des activités de recherche de l'URGI (intégration de données, annotation de répétitions, étude de la structure et de l'évolution des génomes). Elle appartient à l'IFB (Institut Français de Bioinformatique), nœud français du réseau européen de plateformes de bioinformatiques Elixir. Elle a été labellisée par le comité ISC (Infrastructure Scientifique Collective) d'INRAE et par le GIS IBISA (Groupement d'Intérêt Scientifique - Infrastructures en Biologie Sante et Agronomie) comme plateforme stratégique nationale. La plateforme appartient également à l'Infrastructure de Recherche INRAE BioinfOmics et elle fait partie du réseau Science des Plantes de Saclay et la Graduate School Biosphera. Elle est certifiée ISO9001 v. 2015.
Pour plus d'informations, consultez notre site web : https://urgi.versailles.inrae.fr.

Domaines de recherche (selon la classification de l'OCDE) : Biological sciences (Natural sciences), Computer and information sciences
Sources de financement : INRAE - Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement
Partenaires :

- Département INRAE Biologie et Amélioration des Plantes (BAP)
- Département INRAE Ecologie et Biodiversité (ECODIV)

MODES DE GESTION DE LA STRUCTURE

Mode de gestion	Description	Type
Systèmes d'information (SI)	PGD des données gérées dans nos Systèmes d'Information (entrepôt GnpIS et portails de fédération de données)	Dataset
Logiciel	PGD du code source des logiciels que nous développons	Logiciel
Analyse des génomes	PGD des données gérées dans le cadre de nos analyses de génomes	Dataset

CONTRIBUTEURS

Nom, prénom	ORCID	Rôle
Alaux, Michaël	https://orcid.org/0000-0001-9356-4072	Coordinateur du projet
Flores, Raphaël	https://orcid.org/0000-0002-0278-5441	Coordinateur du projet, Contact pour l'infrastructure
Michotey, Célia	https://orcid.org/0000-0003-1877-1703	Responsable du PGD Contact pour les datasets et logiciels SI
Pommier, Cyril	https://orcid.org/0000-0002-9040-8733	Contact pour les datasets et logiciels SI
Confais, Johann	https://orcid.org/0000-0003-2945-5036	Contact pour les datasets et logiciels d'analyse de génomes

HISTORIQUE DES VERSIONS

Date	Version	Statut	Auteur	Validateur
16/02/2021	1	Publié	Célia Michotey	Anne-Françoise Adam-Blondon
04/06/2024	2	Publié	Célia Michotey	Michaël Alaux

INFORMATIONS SUR LA STRUCTURE

NOM DE LA STRUCTURE
Plant Bioinformatics Facility (PlantBioinfoPF)

TYPE DE STRUCTURE
- Plateforme, plateau technique
- ISC (Infrastructure Scientifique Collective)

PlantBioinfoPF fait partie de l'Infrastructure de Recherche BioinfOmics

Fiche ANSci (accès restreint) : https://actif-numerique.inrae.fr/ansci/app/systeme-information/449

IDENTIFIANT DE LA STRUCTURE
https://doi.org/10.15454/1.5572414581735654E12

RESPONSABILITES DANS LA STRUCTURE

Nom, prénom	**Email**	**Rôle**
Alaux, Michaël	michael.alaux@inrae.fr	Responsable scientifique Directeur d'Unité adjoint
Flores, Raphaël	raphael.flores@inrae.fr	Responsable opérationnel
Adam-Blondon, Anne-Françoise	anne-francoise.adam-blondon@inrae.fr	Directrice d'Unité

ETABLISSEMENT(S) TUTELLE(S)
INRAE - Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement

DEPARTEMENT DE RATTACHEMENT INRAE
BAP : Biologie et amélioration des plantes
La plateforme comprend également du personnel ECODIV : Ecologie et Biodiversité (1.8 ETP).

FINANCEUR(S) (PERMETTANT L'ACQUISITION DES JEUX DE DONNEES – HORS PROJETS)
Les jeux de données sont actuellement acquis dans le cadre de projets en collaboration avec des membres de la plateforme. Ils bénéficient de différents types de financements dont les principaux sont : INRAE, ANR (PIA ou non), UE (FP7, H2020, HE).
Nos fournisseurs de données nationaux ou internationaux (e.g. IWGSC) hors projets collaboratifs ne nous fournissent pas leurs sources de financement.

MODE DE GESTION « SYSTEMES D'INFORMATION »

Les systèmes d'information mis à disposition par PlantBioinfoPF sont :

- L'entrepôt de données GnpIS

GnpIS est un système d'information (SI) intégratif et multi-espèces dédié aux plantes et à leurs bioagresseurs. Il permet aux chercheurs d'accéder et de croiser des données de génétiques (accessions, phénotypes, marqueurs, QTL, polymorphismes, génétique d'association) et de génomiques (séquences, cartes physiques, annotations de génome et données d'expression) pour les espèces d'intérêt agronomique et forestier. Le SI est accessible via un portail web et permet de parcourir différents types de données, soit de manière indépendante via des interfaces dédiées, soit de manière simultanée en utilisant des outils de recherche.

- Les portails de fédération de données

Ces portails de recherche, basés sur le même outil, facilitent la découverte et l'accès aux données FAIR (Findable, Accessible, Interoperable, Reusable) à travers une fédération de SI distribués à l'échelle nationale (Portails RARe et BRC4Env, dédiés à la communauté CRB), européenne (FAIDARE, dédié la communauté ELIXIR Plant Sciences) et internationale (WheatIS Data Discovery, dédié à la communauté blé). Ils visent à fournir aux chercheurs un accès simple et rapide aux données biologiques pertinentes en utilisant des mots-clés spécifiques et des filtres.

La gestion des codes des outils liés à ces SI est décrite dans le mode de gestion dédié : « Logiciel ».

PRESENTATION GENERALE DES DONNEES

MODE D'OBTENTION DES DONNEES

- Données produites par un tiers
- Données générées par la structure :
  - Annotation de génome (gènes et éléments transposables issus de prédictions automatiques et/ou de curations)
  - Ontologies de traits

ORIGINE

- Analyse
- Agrégation
- Expérimentation
- Observation

GnpIS stocke et intègre les données de génomique, génétique et phénotypique des chercheurs INRAE et de leurs partenaires nationaux et internationaux.

TYPE DE DONNEES

- Collection
- Dataset
- Logiciel (décrit dans le mode de gestion dédié)

NATURE DES DONNEES

Les données de génomique et génétique de plantes sont fournies par les unités INRAE, leurs partenaires de projets et par le consortium international de séquençage du génome du blé (IWGSC).
GnpIS stocke principalement des données textuelles :

- Ressources génétiques (accessions, données passeport, images)
- Données génomiques (principalement des annotations de génome, du polymorphisme et de la synténie)
- Données génétiques (QTLs et analyse GWAS)
- Données phénotypiques (ontologies, observations et données expérimentales)

FORMAT DES DONNEES

Collections (ressources génétiques)

- CSV/TSV
- XLS/XLSX
- JSON

Datasets

- CSV/TSV
- XLS/XLSX

- VCF
- BED
- Genbank
- EMBL
- GFF
- Fasta/FastQ
- SAM/BAM
- JSON
- RDF

GnpIS a commencé à exposer certaines de ses données dans une représentation sémantique pour améliorer l'intégration de ses données avec d'autres bases de données (format RDF, voir https://urgi.versailles.inra.fr/About-us/News/2017/RDF-Phenotyping).

PERIMETRE THEMATIQUE DES DONNEES

- Biodiversity and Ecology
- Forests and Forest Products
- Insects and Entomology
- Microorganisms
- Omics
- Plant Breeding and Plant Products
- Plant Health and Pathology

DROITS DE PROPRIETE INTELLECTUELLE

QUI DETIENDRA LES DROITS SUR LES DONNEES ET LES AUTRES INFORMATIONS CREEES ?

Notre politique est décrite dans nos conditions d'utilisation. Nous décrivons également les conditions spécifiques à certains jeux de données dans des pages web dédiées.
L'URGI encourage ses utilisateurs à associer des DOI aux jeux de données déposés. Il aide les producteurs de données à obtenir ces DOI et à les associer aux métadonnées appropriées, en collaboration avec le portail pour les données ouvertes Recherche Data Gouv. Cela nous permet de discuter de la licence à associer aux données, avec une proposition de CC-BY V4.0 par défaut pour les données publiques.
Plus généralement, nous implémentons dans GnpIS les termes décrits dans les plans de gestion de données des accords de consortium des projets produisant les données à stocker, certaines données n'étant accessibles qu'à des consortiums d'utilisateurs définis.
La décision de publication des données est prise avec le déposant.

SENSIBILITE DES DONNEES

IDENTIFICATION DU NIVEAU DE SENSIBILITE DES JEUX DE DONNEES

- Public
  La grande majorité des données intégrées dans GnpIS sont des données publiques mises à disposition dans le cadre de la science ouverte. Les données disponibles via les portails fédératifs sont toutes publiques.
- Diffusion limitée
  Certaines données intégrées dans GnpIS sont des données privées car elles sont soumises à des accords de consortium spécifiques, issues de partenaires privés et/ou sous embargo avant publication.
- Confidentiel
  Les données à caractère personnel soumises aux exigences de la CNIL et au RGPD sont confidentielles.
  Sur la plateforme, ces données concernent essentiellement des noms, prénoms et emails.
    - Les informations personnelles fournies lors d'une demande de service ne sont utilisées que par les membres de la plateforme pour traiter cette demande, elles ne sont pas partagées avec des tiers. Ces informations sont conservées dans les meilleures conditions de sécurité et de confidentialité et archivées dans l'outil de gestion de projet de l'URGI (JIRA) pendant la durée de vie des données scientifiques, tel que définie dans le présent PGD.
    - Les informations personnelles fournies lors d'une commande de ressources génétiques (via le panier de commande de FAIDARE, RARe et BRC4Env) ne sont utilisées que par les gestionnaires d'accession des CRB qui vont traiter cette demande. Elles ne sont pas partagées avec des tiers. Ces informations sont conservées dans les meilleures conditions de sécurité et de confidentialité dans une base de données dédiée, tel que définie dans le présent PGD.
    - Les données personnelles incluses dans les jeux de données permettent de conserver la traçabilité des informations. Elles sont donc publiées sous le même statut (public ou privé) que les données

scientifiques associées lors de leur intégration dans GnpIS. Elles seront gérées de la même manière, comme décrit dans ce PGD.

Conformément au règlement européen sur la protection des données personnelles (Règlement européen 2016/679), le propriétaire de données à caractère personnel dispose d'un droit d'accès, de rectification, d'opposition et de suppression des informations le concernant. La plateforme est assistée par le délégué à la protection des données personnelles (DPO) d'INRAE.

QUELLES SONT LES MESURES PRISES ET LES NORMES AUXQUELLES IL EST NECESSAIRE DE SE CONFORMER POUR GARANTIR LA SECURITE DES DONNEES SENSIBLES ?

L'instance de JIRA utilisée par l'URGI pour gérer ses projets, les instances de nos SI (GnpIS, FAIDARE, RARe et BRC4Env) et les bases de données sur lesquels ils reposent sont tous installés sur des serveurs INRAE hébergés dans le data center Ile-de-France.
L'accès à ces outils et aux données sensibles qu'ils contiennent nécessite une authentification régie par le système d'authentification HTTP Apache, les comptes étant créés à la demande par nos services internes.
Dans le cas de GnpIS, un groupe spécifique incluant toutes les personnes qui peuvent accéder à un dataset privé (e.g. collègues, partenaires de projet) est défini avec le propriétaire des données dans le SI. Les données sont étiquetées en base de données avec ce groupe aux droits d'accès définis, de sorte que seules les personnes appartenant au groupe peuvent y accéder. L'authentification sur l'interface web est régie par Apache en suivant les autorisations définies en base de données et décrit ci-dessus.

S'IL Y A DES DONNEES A CARACTERE PERSONNEL, QUELLES SONT LES MESURES ENVISAGEES POUR LES PROTEGER AU COURS DU PROJET OU DANS LE CADRE D'UNE REUTILISATION ?

Les données confidentielles sont gérées de la même façon que les données sensibles : instances des outils utilisés installées sur des serveurs INRAE hébergés dans le data center Ile-de-France, accès aux outils et aux données qu'ils contiennent nécessitant une authentification contrôlée par nos services internes.
Le transfert de données confidentielles à des tiers est soumis à la validation et au suivi du propriétaire des données. Il suit la même procédure que celle décrite ci-dessus, en réutilisant les systèmes d'authentification et d'autorisation sous notre contrôle.

PARTAGE DES DONNEES

Y A T'IL UNE OBLIGATION DE PARTAGE (OU A L'INVERSE UNE INTERDICTION OU UNE RESTRICTION) ?

Le fournisseur de données s'engage à ouvrir publiquement les données (open data). Une période d'embargo peut être définie par rapport à l'usage de la communauté scientifique.
La décision de publication des données est prise avec le responsable scientifique au moment de la soumission.
Toutefois, la plateforme doit être citée dans toute publication mentionnant le(s) jeu(x) de données intégré(s) : *"This work was performed with the facilities of the Plant Bioinformatics Facility (https://doi.org/10.15454/1.5572414581735654E12)".*

QUELLES SONT LES REUTILISATIONS POTENTIELLES DE CES DONNEES ?

- Mise à jour des catalogues internationaux de ressources génétiques
- Mise à jour d'autres systèmes d'information
- Réutilisation pour de nouvelles fins de recherche
- Soutien aux politiques publiques, expertise
- Formations

LA LECTURE DES DONNEES NECESSITE-T-ELLE LE RECOURS A UN LOGICIEL OU UN OUTIL SPECIFIQUE ? SI OUI, LEQUEL ?

Les données sont accessibles via les interfaces web des SI (GnpIS, FAIDARE, WheatIS search, RARe & BRC4Env) et via des services web standardisés (API RESTful) qui permettent un accès automatique par programmation.
Les données peuvent être téléchargées dans différents formats texte standards (CSV/TSV, GFF, VCF, JSON ...).

COMMENT LES DONNEES SERONT-ELLES PARTAGEES ?

Les données sont partagées via les interfaces web et les API de services web mis à disposition par les SI. Elles peuvent être téléchargées dans différents formats texte standards (CSV/TSV, GFF, VCF, JSON ...).
Une authentification est nécessaire pour accéder aux données privées (voir nos conditions d'utilisation).
Certains jeux de données sont également disponibles dans le dataverse URGI du portail Recherche Data Gouv.

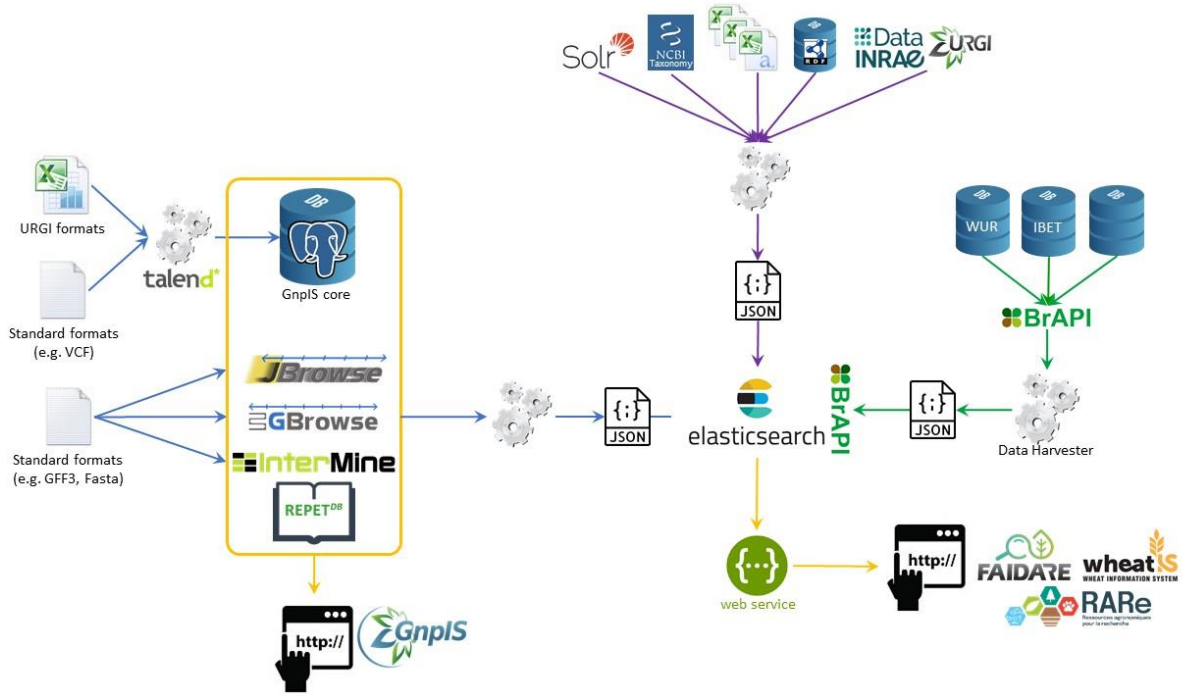
AVEC QUI ?

- Les données publiques sont en accès libre donc partagées avec n'importe qui
- Les données privées sont partagées avec des utilisateurs identifiés (partenaires académiques et/ou privés)

SOUS QUELLE LICENCE ?
Nous encourageons et nous accompagnons les responsables scientifiques à associer un DOI à leurs jeux de données via une publication dans le dataverse URGI du portail Recherche Data Gouv. Cela nous permet de discuter de la licence à associer aux données.
Par défaut, les données publiques sont sous licence CC-BY V4.0.

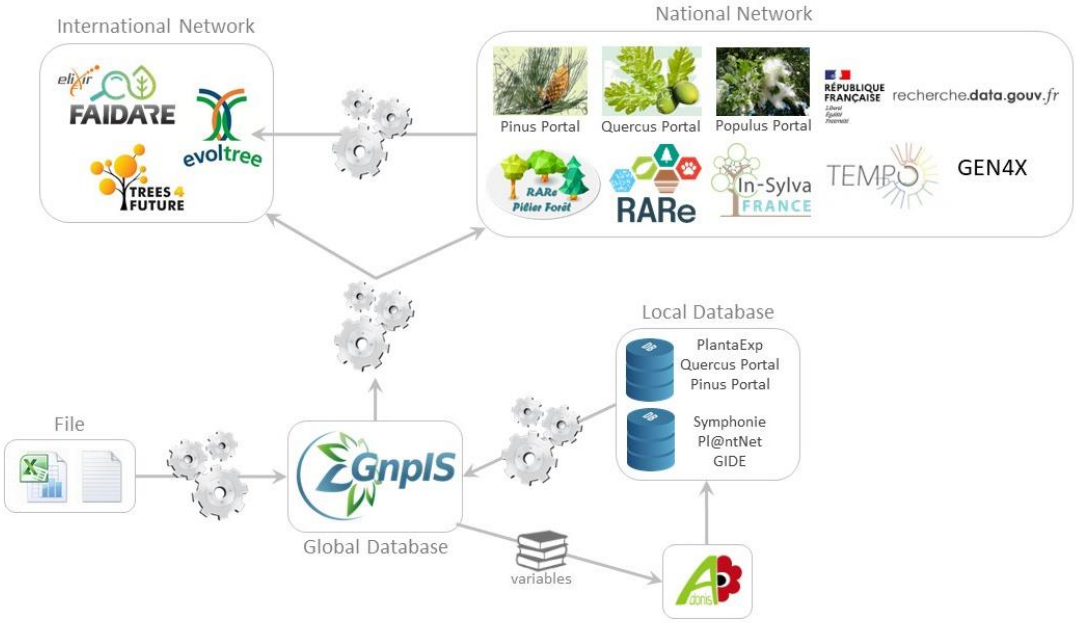
ORGANISATION ET DOCUMENTATION DES DONNEES

QUELS METHODES ET OUTILS SONT UTILISES POUR ACQUERIR ET TRAITER LES DONNEES, DEPUIS LEUR ACQUISITION JUSQU'A LEUR MISE A DISPOSITION, LEUR ARCHIVAGE OU LEUR DESTRUCTION ?
Des procédures qualité (norme ISO9001 v. 2015, accès restreint) existent pour expliquer en détail comment gérer les données dans les différents SI de la plateforme (acquisition, validation, intégration et visualisation dans le SI, archivage). Le schéma ci-dessous illustre cette gestion des données (https://urgi.versailles.inrae.fr/files/gnpis/data_integration_gnpis.jpg).



Les données produites par les équipes de recherche sont soumises à PlantBioinfoPF en utilisant des formats d'échange standards et internes. Des outils ETL - Extract Transform Load - (principalement des projets Talend Open Studio et des scripts Perl et Python) sont utilisés pour vérifier la qualité et la cohérence des données et pour alimenter les bases de données de GnpIS (PostgreSQL et MySQL). Les données provenant de GnpIS et des autres SI à afficher dans les portails fédératifs sont intégrées ensemble dans une base de données Elasticsearch (NoSQL orienté document) à l'aide d'outils ETL (principalement des scripts Bash et Python). Les données sont ensuite rendues accessibles via des API RESTful et des interfaces web (GnpIS, FAIDARE, WheatIS search, RARe & BRC4Env, Siregal).

Dans le cas des arbres forestiers, différents types de données sont gérés dans divers SI avec une accessibilité différente. Pour partager ces connaissances, nous avons mis en place un flux de données automatisé pour synchroniser les données partagées par plusieurs SI, comme le montre le schéma ci-dessous (https://urgi.versailles.inrae.fr/files/gnpis/forest_interoperability.jpg).



Les données produites par les équipes de recherche sont gérées dans les SI locaux : outils d'acquisition de données, tel qu'Adonis, bases de données locales et fichiers (formats standards ou internes). Lorsque les données sont prêtes à être partagées, elles sont extraites de ces sources locales et insérées dans GnpIS, notre SI global. Si nécessaire, les données peuvent ensuite être extraites de GnpIS pour être insérées dans d'autres SI afin d'améliorer leur visibilité à l'échelle nationale et/ou internationale. Des outils ETL spécifiques à chaque SI sont utilisés pour gérer automatiquement ces flux de données.

QUELLES METADONNE SERONT UTILISEES POUR ACCOMPAGNER LE JEU DE DONNEES ?
QUELS SERONT LES STANDARDS, VOCABULAIRES, TAXONOMIES... UTILISES POUR DECRIRE ET REPRESENTER LES DONNEES ET ELEMENTS DE METADONNEES ?
COMMENT LES METADONNEES SERONT-ELLES PRODUITES ET MISES A JOUR ?

Métadonnées	**Origine des métadonnées, mode de production**	**Standard et vocabulaire associé**	**Conditions et/ou fréquence de mise à jour**
Données passeport des ressources génétiques	Alimentation du format de soumission de GnpIS (semi-manuel) par les gestionnaires des ressources	• Multi-Crop Passport Descriptors (MCPD) • Taxonomie de référence (NCBI, TAXREF, Catalogue of Life …) • Ontologie de traits (CropOntology Trait dictionary, Trait Ontology …)	• Le MCPD est sous la gouvernance de la FAO • Les mises à jour du Trait dictionary dépendent de la communauté CropOntology
Expériences de phénotypage des plantes	Alimentation du format de soumission de GnpIS (semi-manuel) par les producteurs de données	• Minimal Informations About Plant Phenotyping Experiments (MIAPPE) • Breeding API (BrAPI) • CropOntology Trait dictionary	• www.miappe.org • https://brapi.org/ • Les mises à jour du Trait dictionary dépendent de la communauté
Données génomiques	Alimentation du format de soumission de GnpIS (semi-manuel) par les producteurs de données	Recommandations EMBL, NCBI	

UNE DOCUMENTATION COMPLEMENTAIRE AUX METADONNEES EST-ELLE NECESSAIRE POUR DECRIRE LES DONNEES ET ASSURER LEUR REUTILISABILITE SUR LE LONG TERME ?

Les formats de soumission de GnpIS garantissent que les métadonnées essentielles pour décrire et réutiliser les jeux de données sont disponibles. Si nécessaire, des documents peuvent être joints aux données pour fournir des informations supplémentaires (par exemple, des fichiers readme), les formats non propriétaires étant fortement recommandés. En outre, nous associons de plus en plus souvent des DOI aux jeux de données et nous recommandons la publication de « data papers » à nos utilisateurs afin d'améliorer cette documentation.
Les données peuvent également être récupérées dans des formats standard :

- Les ontologies de traits peuvent être récupérées sous différents formats suivant le standard du CropOntology Trait dictionary (Excel, CSV, JSON) via l'interface web de GnpIS ou via un service web compatible avec la BrAPI (Breeding API, une API RESTful standardisée qui permet l'interopérabilité entre les bases de données de sélection végétale).
- Les accessions et leurs données passeport peuvent être récupérées dans un format CSV conforme au MCPD (Multi-Crop Passport Descriptors, un standard qui facilite l'échange d'informations sur le matériel génétique) via l'interface web de GnpIS ou un service web compatible avec la BrAPI.
- Les données de phénotypage peuvent être récupérées dans un format ISA-Tab conforme au MIAPPE (Minimal Informations About Plant Phenotyping Experiments, un standard spécifiant les métadonnées nécessaires pour décrire correctement les expériences de phénotypage des plantes) de manière informatique via un service web compatible avec la BrAPI.
- Les données de polymorphisme peuvent être récupérées aux formats VCF et CSV via l'interface web de GnpIS.
- Les données d'annotation du génome peuvent être récupérées au format GFF3 via des pages web dédiées à l'espèce lorsque GnpIS est le dépôt officiel ou dans des archives internationales comme décrit dans les métadonnées des instances JBrowse de GnpIS.
- Les données génétiques peuvent être récupérées au format CSV car il n'existe pas de standard international.

Il existe des outils pour produire ces métadonnées :

- Liste d'outils de métadonnées du Digital Curation Center
- Répertoire de métadonnées de la RDA (Research Data Alliance)

COMMENT LES FICHIERS DE DONNEES SONT-ILS GERES ET ORGANISES : CONTROLE DES VERSIONS, CONVENTIONS DE NOMMAGE DES FICHIERS, ORGANISATION DES FICHIERS

Des procédures qualité (accès restreint) existent pour définir en détail comment gérer les données dans nos SI (GnpIS et les portails fédératifs).
Les outils ETL (scripts Bash, Perl ou Python et projets Talend Open Studio) sont utilisés pour traiter les fichiers de données et contrôler la qualité et l'intégrité des données tout au long de leur gestion, de la soumission à l'accès via le portail GnpIS. L'implémentation de *workflow* ETL reproductibles est en cours de développement.
Les ontologies de traits que nous développons et maintenons sont gérées dans un logiciel de gestion de version, plus précisément sur un projet public de la forgemia qui est la forge GitLab du département MATHNUM d'INRAE (voir Ontologies).

QUEL EST LE PROCESSUS DE CONTROLE QUALITE DES DONNEES ?

Des procédures qualité (accès restreint) existent pour définir en détail comment gérer les données dans nos SI (GnpIS et les portails fédératifs), y compris les étapes pour contrôler la qualité et l'intégrité des données.
Les étapes permettant de vérifier la validité des fichiers soumis (e.g. format et contenu obligatoire respectés), la qualité des données (e.g. vocabulaires/ontologies contrôlés correctement utilisés) et la cohérence du jeu de données (e.g. cohérence des données et des liens entre elles) sont implémentées dans les outils ETL (scripts Bash, Perl ou Python et projets Talend Open Studio) que nous utilisons pour gérer les jeux de données dans nos SI. De plus, l'utilisation de bases de données relationnelles pour GnpIS (SGBDR PostgreSQL et MySQL) garantit l'intégrité des données en s'appuyant sur certaines contraintes métier mises en place au fil des années. Enfin, lorsqu'une insertion est effectuée, plusieurs vérifications concernant le nombre d'entrées sont également effectuées afin de s'assurer qu'aucune donnée n'est manquante ou dupliquée.

STOCKAGE ET SECURITE DES DONNEES

LES SYSTEMES D'INFORMATION DE LA STRUCTURE ONT-ILS FAIT L'OBJET D'UNE ANALYSE DE RISQUES OU D'UNE HOMOLOGATION ?

Nos SI ne sont pas encore homologués, mais une démarche d'analyse de risques est en cours dans le cadre de la mise en place de la politique de sécurité des SI d'INRAE.

QUELS TYPES DE SUPPORTS PHYSIQUES SONT UTILISES POUR STOCKER LES DONNEES ?
Les producteurs de données peuvent accéder au service de dépôt de données de différentes manières, qui sont décrites sur le site web de l'URGI :

- Via l'offre de service de la plateforme
- Via le formulaire de soumission de GnpIS

Et des pages spécifiques dédiées à la communauté blé :

- Le groupe d'expert WheatIS de la Wheat Initiative
- L'entrepôt de données de l'IWGSC

Les fichiers soumis sont stockés sur un volume NetApp/CEPH monté sur les serveurs de l'URGI. Le contenu des fichiers est traité pour que les données soient insérées dans les bases de données composant nos SI (SGBDR MySQL et PostgreSQL, NoSQL Elasticsearch). Ces bases de données sont hébergées sur des machines virtuelles (VM) dans nos instances de virtualisation (Proxmox et cloud OpenStack). Le tout est hébergé dans le data center Ile-de-France d'INRAE.
Les ontologies de traits que nous développons et maintenons sont versionnées sur un projet public de la forge GIT forgemia (voir Ontologies).

QUELLES SONT LES MESURES DE SECURITE MISES EN PLACE LORS DES ETAPES DE TRANSFERT DES DONNEES ?
Les données téléchargeables depuis nos SI (interfaces et services web) sont transférées par le réseau via le protocole chiffré HTTPS. Si les données sont privées, une authentification régie par le système d'authentification HTTP Apache et sous notre contrôle est nécessaire avant de pouvoir lancer le téléchargement.
Les données versionnées dans le GIT de la forgemia sont récupérables via les protocoles sécurisés HTTPS ou SSH.

QUELLE EST LA VOLUMETRIE ACTUELLE ET PREVISIONNELLE ?
Jeux de données : 6.5 TB de stockage en Juin 2024
Ontologies de traits : 116 MB en Juin 2024

L'ENTITE HEBERGEANT PHYSIQUEMENT LES DONNEES A-T-ELLE UNE POLITIQUE DE SECURITE DE L'INFORMATION ET A-T-ELLE UN PLAN D'ASSURANCE SECURITE ?
Notre politique de sécurité suit la charte des infrastructures de recherche d'INRAE.
Toutes les données soumises et les fichiers joints sont stockés sur les serveurs de l'URGI, hébergés dans le data center Ile-de-France d'INRAE, et régulièrement sauvegardés. Les bases de données composant nos SI sont sauvegardées via des dumps deux fois par mois avec une rétention de 2 mois. Les volumes de stockage NetApp sont sauvegardés par snapshot deux fois par jour. Ces sauvegardes sont dupliquées mensuellement sur le data center INRAE de Toulouse.

SECURITE - CONFIDENTIALITE : LES DONNEES FONT-ELLES L'OBJET D'ECHANGE OU DE PARTAGE AVEC DE TIERS ACTEURS ET SELON QUELLES MODALITES ? COMMENT SONT DETERMINES LES DROITS D'ACCES AUX DONNEES AVANT LEUR PUBLICATION ?
Les données privées ne sont accessibles qu'après un processus d'authentification. Les droits d'accès sont donnés à un groupe spécifique comprenant toutes les personnes qui peuvent accéder aux données (e.g. collègues, partenaires de projet) et définis avec le propriétaire des données. Les données sont étiquetées dans GnpIS avec ce groupe, de sorte que seules les personnes appartenant à ce groupe peuvent y accéder.
Le transfert de données privées à des tiers est soumis à la validation et au suivi du propriétaire des données.

SECURITE - INTEGRITE - TRACABILITE : QUELLES SONT LES MESURES DE PROTECTION MISES EN ŒUVRE POUR SUIVRE LA PRODUCTION ET L'ANALYSE DES DONNEES ?
Les données privées peuvent être fournies à l'aide d'outils cryptographiques, telles que des sessions shell sécurisées pour copier les données sur nos serveurs qui sont soutenus par la technologie NetApp/CEPH, qui est configurée pour prendre un instantané de toutes les données copiées deux fois par jour. L'accès aux fichiers copiés est contrôlé à l'aide de permissions UNIX sur des serveurs Linux qui montent les volumes NetApp/CEPH uniquement lorsque c'est nécessaire.
Les fichiers soumis suivent les procédures qualité (accès restreint) déjà mentionnées et toutes les étapes sont enregistrées dans notre instance interne de JIRA (cahier de laboratoire numérique) dans une tâche dédiée.
Les ontologies de traits sont gérées avec les fonctionnalités d'une forge GIT (forgemia).

LES AGENTS DE LA STRUCTURE ONT-ILS BENEFICIE D'UNE SENSIBILISATION AUX BONNES PRATIQUES D'HYGIENE NUMERIQUE ?
Oui.
Tous les agents ont reçu les communications de la cellule SSI INRAE. Certains d'entre eux, en première ligne, ont également suivi une formation à la sécurité informatique offensive et défensive et sont en lien avec la chaîne SSI INRAE.

ARCHIVAGE ET CONSERVATION DES DONNEES

QUELLES SONT LES DONNEES A CONSERVER SUR LE MOYEN OU LE LONG TERME ET QUELLES SONT LES DONNEES A DETRUIRE ?

Les données à conserver sur le long terme sont :

- Les données passeport des ressources génétiques et les images associées
- Les données de phénotypage traitées (pas les fichiers de données brutes des capteurs)
- Les ontologies de traits que nous gérons
- Les données d'annotation des éléments transposables

SUR QUELLE PLATEFORME D'ARCHIVAGE PERENNE SERONT ARCHIVEES LES DONNEES A CONSERVER SUR LE LONG TERME ? SINON, QUELLES PROCEDURES SERONT MISES EN PLACE POUR LA CONSERVATION A LONG TERME ?

Les ontologies de traits sont versionnées sur un dépôt public du GitLab de la forgemia (voir forgemia Ontologies). Ce dépôt a été synchronisé avec Software Heritage (une initiative pour préserver et partager les codes source publiques) pour permettre l'archivage automatiquement des données qu'il contient à intervals réguliers (voir software heritage Ontologies).

Sinon, il n'y a pas d'action d'archivage au-delà de GnpIS lui-même. Cependant, nous imposons progressivement la publication des jeux de données en parallèle dans le dataverse URGI du portail Recherche Data Gouv, ce qui, en association avec le DOI, garantit un accès aux données pendant 10 ans.

QUELLE EST LA DUREE DE CONSERVATION DES DONNEES ?

Les ontologies de traits seront accessibles au public aussi longtemps que la forgemia (ou un équivalent) existera. Une solution alternative sera utilisée dans le cas improbable de l'arrêt de ces solutions.

Les ressources génétiques et les données de phénotypage sont stratégiques et elles seront conservées aussi longtemps qu'INRAE en fournira les moyens.

Les annotations TE sont également versionnées et conservées aussi longtemps que possible, car il n'existe pas de dépôt central en libre accès.

Pour les autres données, cela dépend de l'intérêt d'être intégré dans le GnpIS avec d'autres jeux de données et de questions stratégiques à discuter avec le fournisseur de données.

QUELLES GARANTIES DE FINANCEMENTS COUVRIRONT LES COUTS ASSOCIES A LA CONSERVATION A LONG TERME ?

Les SI de la plateforme sont des actifs stratégiques d'INRAE, qui fournit des financements de base (ressources humaines et soutien financier de la plateforme) via les départements de recherche BAP et ECODIV ainsi que la CNOC (Commission Nationale des Outils Collectifs).

La plateforme a également accès à des financements dans le cadre de projets qui permettent de soutenir ces SI.

MODE DE GESTION « LOGICIEL »

Les logiciels développés et mis à disposition par PlantBioinfoPF sont :

- Des outils liés à nos systèmes information (SI)
    - L'entrepôt de données GnpIS et les portails de fédération de données FAIDARE, WheatIS Data Discovery, RARe et BRC4Env
    - Les ETL (Extract Transform Load) à utiliser pour alimenter les SI en données
    - RARe-Basket (accès restreint), un outil de gestion pour la commande de ressources génétiques
    - Trait ontology widget, un outil de visualisation d'ontologies compatibles avec le standard de la CropOntology
- Des outils d'analyse de génome, comme la suite REPET et TE_finder pour l'annotation des éléments transposables, Caulifinder pour l'annotation des éléments viraux endogènes (EVE) d'origine caulimoviride.

La gestion des données dans ces outils est décrite dans les modes de gestion dédiés : « Systèmes d'information » et « Analyse des génomes ».

PRESENTATION GENERALE DES DONNEES

MODE D'OBTENTION DES DONNEES

- Données produites par un tiers
- Données générées par la structure

Le code de nos logiciels étant ouvert, des tiers peuvent y contribuer et ajouter de nouvelles fonctionnalités.

ORIGINE

- Code

TYPE DE DONNEES

- Logiciel
- Workflow
- Service

NATURE DES DONNEES

Les logiciels sont les outils que nous développons :

- Systèmes d'information implémentés en Java pour le backend et Angular ou le framework GWT pour le frontend. La couche donnée s'appuie sur des bases de données relationnelles PostgreSQL et MySQL ainsi qu'un cluster Elasticsearch (NoSQL orienté document).
- Outils d'analyse de génome implémentés en Python et C++.

Les workflows sont :

- Outils ETL utilisés pour alimenter nos SI en données. Les scripts sont développés en Bash, Perl ou Python et des projets Talend Open Studio.
- Pipelines d'outils d'analyses, telle que la suite REPET pour l'annotation des éléments transposables, développées avec SnakeMake (gestionnaire de workflow) en Python.

Les services sont des services web RESTful, par exemple notre implémentation de la Breeding API - BrAPI, développés en Java.

FORMAT DES DONNEES

Logiciel

- JSON
- Python
- C++
- Java
- TypeScript
- HTML
- SCSS
- YAML
- Markdown

Workflow

- JSON
- SQL
- Bash

- Perl
- Python
- C++
- Talend Open Studio
- SnakeMake
- Ansible
- DockerFile

PERIMETRE THEMATIQUE DES DONNEES

- Information management
- Omics

DROITS DE PROPRIETE INTELLECTUELLE

QUI DETIENDRA LES DROITS SUR LES DONNEES ET LES AUTRES INFORMATIONS CREEES ?

Les institutions qui ont payé les développements (principalement INRAE) sont propriétaires du code.

SENSIBILITE DES DONNEES

IDENTIFICATION DU NIVEAU DE SENSIBILITE DES JEUX DE DONNEES

- Public
  A part pour GnpIS (cf. ci-dessous) tous nos codes sont ouverts et mis à disposition dans un logiciel de gestion de version. La grande majorité de ces codes sont accessibles publiquement mais les branches principales de leurs dépôts sont restreintes pour certaines actions. Les autres sont disponibles uniquement sur demande.
- Diffusion limitée
  Le modèle de la base de données historique du SI GnpIS, GnpIS-coreDB, est privé. Tous les codes liés à la structure de cette base de données sont donc à diffusion limitée (backend du SI, ETL).

QUELLES SONT LES MESURES PRISES ET LES NORMES AUXQUELLES IL EST NECESSAIRE DE SE CONFORMER POUR GARANTIR LA SECURITE DES DONNEES SENSIBLES ?

Le code de nos logiciels et outils ETL est géré dans un logiciel de gestion de version, plus précisément sur des projets de la forgemia qui est la forge GitLab du département MATHNUM d'INRAE. Nous utilisons les fonctionnalités de l'application pour gérer l'accès et les droits sur nos dépôts.

S'IL Y A DES DONNEES A CARACTERE PERSONNEL, QUELLES SONT LES MESURES ENVISAGEES POUR LES PROTEGER AU COURS DU PROJET OU DANS LE CADRE D'UNE REUTILISATION ?

Nos codes ne contiennent pas de données à caractère personnel.

PARTAGE DES DONNEES

Y A T'IL UNE OBLIGATION DE PARTAGE (OU A L'INVERSE UNE INTERDICTION OU UNE RESTRICTION) ?

Il n'y a aucune obligation de partage de nos codes, mais nous encourageons les utilisateurs à s'engager (e.g. en nous faisant des retours, en identifiant les bogues) et les développeurs à améliorer le code (e.g. en corrigeant les bogues, en ajoutant de nouvelles fonctionnalités) afin d'enrichir le dépôt GIT original pour que tout le monde puisse en profiter. Cependant, le partage du modèle de la base de données historique de GnpIS, GnpIS-coreDB, est interdit.

QUELLES SONT LES REUTILISATIONS POTENTIELLES DE CES DONNEES ?

- Création d'une nouvelle instance dédiée à une communauté spécifique
- Amélioration de fonctionnalités existantes
- Ajout de nouvelles fonctionnalités pour répondre à de nouveaux besoins

LA LECTURE DES DONNEES NECESSITE-T-ELLE LE RECOURS A UN LOGICIEL OU UN OUTIL SPECIFIQUE ? SI OUI, LEQUEL ?

Les outils ETL développés avec Talend Open Studio doivent être gérés avec l'outil dédié (Talend Open Studio for Data Integration, accessible gratuitement).
Sinon, aucun logiciel ou outil spécifique n'est nécessaire pour lire le code.

COMMENT LES DONNEES SERONT-ELLES PARTAGEES ?

- Le code de GnpIS et de ses ETL est versionné sur des dépôts privés de la forge GIT forgemia et sont disponibles uniquement sur demande.

- Le code des portails fédératifs et de leurs ETL est versionné sur des dépôts publics de la forge GIT forgemia (voir Data Discovery, FAIDARE, ETL_data_portals et RARe-Basket). Le dépôt FAIDARE est également mis en miroir sur GitHub (voir FAIDARE).
- Le Trait Ontology widget et l'ETL utilisé pour alimenter le portail FAIDARE en données sont disponibles publiquement sur GitHub (voir Trait Ontology widget et ETL FAIDARE).
- Les outils d'analyse de génome sont versionnés sur des dépôts de la forge GIT forgemia (voir repet_pipe, TE_finder et Caulifinder). Certains codes sont également versionnés sur des dépôts publics sur GitHub (voir TE_finder).

AVEC QUI ?

- Nos ETL et logiciels sont partagés avec tout le monde (accès ouvert).
- Le modèle de la base de données historique de GnpIS, GnpIS-coreDB, est partagé suivant la licence propriétaire.

SOUS QUELLE LICENCE ?

- Le code de nos portails fédératifs et des ETL utilisés pour les alimenter en données est sous licence BSD 3-Clause : la redistribution et l'utilisation sous forme source et binaire, avec ou sans modification, sont autorisées sous certaines conditions.
- Les ETL de GnpIS sont sous licence GNU LGPL v3 : tout le monde est autorisé à copier et distribuer des copies verbatim, mais il est interdit de les modifier.
- Le modèle de la base de données historique de GnpIS, GnpIS-coreDB, est sous licence propriétaire et protégé par des dépôts auprès de l'Agence de Protection des Programmes.
- Les outils REPET sont sous licence CeCILL v2.1 : accorde aux utilisateurs le droit de copier, modifier et distribuer le logiciel régi par cette licence sous un modèle de distribution open source.
- Caulifinder est sous licence MIT : exige uniquement la préservation des mentions de copyright et de licence.

ORGANISATION ET DOCUMENTATION DES DONNEES

QUELS METHODES ET OUTILS SONT UTILISES POUR ACQUERIR ET TRAITER LES DONNEES, DEPUIS LEUR ACQUISITION JUSQU'A LEUR MISE A DISPOSITION, LEUR ARCHIVAGE OU LEUR DESTRUCTION ?

Nous utilisons les fonctionnalités du GitLab de la forgemia pour gérer nos codes.
A l'exception des ETLs Talend Open Studio qui sont versionnés tels quels dans la branche principale du dépôt (master), toute modification de code entraine la création d'une branche secondaire dédiée (feature). La branche feature est créée à partir de la branche master et les modifications à apporter au code y sont versionnées (commit) jusqu'à ce que le code soit prêt à passer production. A ce moment-là, la branche feature est fusionnée dans la branche master via une demande de fusion (merge request) qui implique une revue de code et le succès de la réalisation d'une batterie de tests.

QUELLES METADONNE SERONT UTILISEES POUR ACCOMPAGNER LE JEU DE DONNEES ?
QUELS SERONT LES STANDARDS, VOCABULAIRES, TAXONOMIES... UTILISES POUR DECRIRE ET REPRESENTER LES DONNEES ET ELEMENTS DE METADONNEES ?
COMMENT LES METADONNEES SERONT-ELLES PRODUITES ET MISES A JOUR ?

- Le modèle de données utilisé dans Data Discovery (le code générique qui sert de base aux portails fédératifs) est dérivé du modèle générique défini collectivement dans Spannagl et al. 2016.
- FAIDARE s'appuie sur les spécifications de la Breeding API (BrAPI), elle-même s'appuyant sur les standards Minimal Informations About Plant Phenotyping Experiments (MIAPPE) et Multi-Crop Passport Descriptors (MCPD). Sa mise à jour suit le processus mis en place par la communauté BrAPI.
- Le widget Trait Ontology est basé sur les web services « observation variable » de la BrAPI et sa mise à jour suit le processus mis en place par la communauté BrAPI.

UNE DOCUMENTATION COMPLEMENTAIRE AUX METADONNEES EST-ELLE NECESSAIRE POUR DECRIRE LES DONNEES ET ASSURER LEUR REUTILISABILITE SUR LE LONG TERME ?

Des fichiers README sont disponibles dans chaque dépôt de logiciel pour expliquer comment contribuer, installer les prérequis, construire et exécuter l'application, installer l'intégration continue (IC) et mettre en place la configuration et l'authentification. Des tutoriels expliquant comment utiliser nos outils d'analyse et une documentation expliquant comment rejoindre nos portails fédératifs (voir Comment rejoindre la fédération Data-Discovery) sont également disponibles avec le code.
Pour les outils ETL, une combinaison de fichiers README versionnés dans le dépôt de code et de guides de bonnes pratiques internes stockés sur notre Wiki ou SharePoint (accès restreint) sont disponibles.

COMMENT LES FICHIERS DE DONNEES SONT-ILS GERES ET ORGANISES : CONTROLE DES VERSIONS, CONVENTIONS DE NOMMAGE DES FICHIERS, ORGANISATION DES FICHIERS

Le Trait Ontology widget et l'ETL utilisé pour alimenter le portail FAIDARE en données sont versionnés sur GitHub (voir ETL FAIDARE et widget Trait Ontology). Les autres codes sont versionnés sur le GitLab de la forgemia.
Des fichiers expliquant comment contribuer au code sont également disponibles aux côtés du code. Ils décrivent comment gérer le code dans GIT, comment gérer les données de l'application, comment installer l'environnement de développement et faire des tests (voir Contribuer à Data-Discovery).

QUEL EST LE PROCESSUS DE CONTROLE QUALITE DES DONNEES ?

Comme expliqué dans les fichiers de contribution, chaque modification de code (e.g. correction de bogues, nouvelle fonctionnalité, changement de version) doit être effectuée dans une nouvelle branche dédiée du dépôt GIT. Lorsque la branche est prête à être fusionnée dans la branche principale du dépôt, une demande de fusion doit être créée. Au moins un *core committer* de notre équipe revoit le code avant validation et une série de tests est lancée et doit réussir pour que la branche soit fusionnée et le code passé en production.
Pour nos logiciels qui sont mis en production via une livraison continue, c'est-à-dire que chaque modification de la branche master entraine une mise en production du code (cas de FAIDARE, WheatIS Data Discovery, RARe et BRC4Env, RARe-Basket), une intégration continue est automatiquement lancée par GitLab à chaque commit ou création de merge-request (voir le fichier .gitlab-ci.yml du dépôts Data Discovery).

STOCKAGE ET SECURITE DES DONNEES

LES SYSTEMES D'INFORMATION DE LA STRUCTURE ONT-ILS FAIT L'OBJET D'UNE ANALYSE DE RISQUES OU D'UNE HOMOLOGATION ?

Nos SI ne sont pas encore homologués, mais une démarche d'analyse de risques est en cours dans le cadre de la mise en place de la politique de sécurité des SI d'INRAE.

QUELS TYPES DE SUPPORTS PHYSIQUES SONT UTILISES POUR STOCKER LES DONNEES ?

Nos codes sont versionnés sur des forges GIT : GitLab de la forgemia et/ou GitHub.

QUELLES SONT LES MESURES DE SECURITE MISES EN PLACE LORS DES ETAPES DE TRANSFERT DES DONNÉES ?

Les données versionnées dans le GitLab forgemia sont récupérables via les protocoles sécurisés HTTPS ou SSH. Pour GitHub le transfert se fait via le protocole sécurisé HTTPS.

QUELLE EST LA VOLUMETRIE ACTUELLE ET PREVISIONNELLE ?

- Système d'information : 12,5 GB
- ETLs : 1 GB
- Outils d'analyse de génome : 10 GB

L'ENTITE HEBERGEANT PHYSIQUEMENT LES DONNEES A-T-ELLE UNE POLITIQUE DE SECURITE DE L'INFORMATION ET A-T-ELLE UN PLAN D'ASSURANCE SECURITE ?

La forgemia est installée sur une VM hébergée dans le data center INRAE de Toulouse et elle va évoluer pour devenir la forge institutionnelle d'INRAE.
Afin de garantir une haute disponibilité et robustesse, le service est virtualisé sur un serveur hébergé dans le data center "L'ARCHE DE DONNÉES Francis Sevila" situé sur le centre Occitanie-Toulouse d'INRAE. L'administration du serveur, les sauvegardes quotidiennes et la gestion de GitLab sont assurées par une équipe d'administrateurs système provenant de différentes unités du département MATHNUM d'INRAE en collaboration avec l'équipe informatique du centre de Toulouse.

SECURITE - CONFIDENTIALITE : LES DONNEES FONT-ELLES L'OBJET D'ECHANGE OU DE PARTAGE AVEC DE TIERS ACTEURS ET SELON QUELLES MODALITES ? COMMENT SONT DETERMINES LES DROITS D'ACCES AUX DONNEES AVANT LEUR PUBLICATION ?

Les codes versionnés sur des dépôts publics sont en accès libre.
Les codes versionnés sur des dépôts privés sont accessibles via une authentification LDAP. La gestion des accès et des droits sur ces dépôts est sous notre contrôle grâce à l'utilisation des fonctionnalités offertes par GitLab.

SECURITE - INTEGRITE - TRACABILITE : QUELLES SONT LES MESURES DE PROTECTION MISES EN ŒUVRE POUR SUIVRE LA PRODUCTION ET L'ANALYSE DES DONNEES ?

Nous utilisons les fonctionnalités d'une forge GIT (GitLab de la forgemia et/ou GitHub) pour gérer nos codes.

LES AGENTS DE LA STRUCTURE ONT-ILS BENEFICIE D'UNE SENSIBILISATION AUX BONNES PRATIQUES D'HYGIENE NUMERIQUE ?
Oui.
Tous les agents ont reçu les communications de la cellule SSI INRAE. Certains d'entre eux, en première ligne, ont également suivi une formation à la sécurité informatique offensive et défensive et sont en lien avec la chaîne SSI INRAE.

ARCHIVAGE ET CONSERVATION DES DONNEES

QUELLES SONT LES DONNEES A CONSERVER SUR LE MOYEN OU LE LONG TERME ET QUELLES SONT LES DONNEES A DETRUIRE ?
Nos codes sont à conserver sur le long terme, ils ne seront pas détruits.

SUR QUELLE PLATEFORME D'ARCHIVAGE PERENNE SERONT ARCHIVEES LES DONNEES A CONSERVER SUR LE LONG TERME ?
QUELLES PROCEDURES SERONT MISES EN PLACE POUR LA CONSERVATION A LONG TERME ?
Les dépôts GitLab publiques ont été synchronisés avec Software Heritage (une initiative pour préserver et partager les codes source publiques) pour permettre l'archivage automatique de nos codes à intervalles réguliers (voir Data-Discovery, FAIDARE, ETL_data_portals, RARe-Basket, TE_finder et event_caulifinder).
Les dépôts GitHub publiques ont été archivés sur Software Heritage manuellement et une seule fois (voir Trait Ontology widget, ETL FAIDARE). Un mécanisme plus pérenne est à mettre en place (sous le même principe que le système de Gitlab).

QUELLE EST LA DUREE DE CONSERVATION DES DONNEES ?
Comme nos codes sont versionnés sur des forges GIT (GitLab de la forgemia et/ou GitHub) ils seront accessibles tant que ces forges existeront. Une solution alternative sera utilisée dans le cas improbable de l'arrêt de ces solutions.

QUELLES GARANTIES DE FINANCEMENTS COUVRIRONT LES COUTS ASSOCIES A LA CONSERVATION A LONG TERME ?
NA. Le service de mise à disposition de codes offert par la plateforme est gratuit. Les coûts de la forgemia sont actuellement supportés par le département MATHNUM d'INRAE et ils seront directement soutenus par INRAE quand ce GIT deviendra la forge institutionnelle.

MODE DE GESTION « ANALYSE DES GENOMES »

PRESENTATION GENERALE DES DONNEES

MODE D'OBTENTION DES DONNEES

- Données produites par un tiers
- Données générées par la structure : Annotations de génome (gènes et éléments mobiles issus de prédictions automatiques et/ou de curations)

ORIGINE

- Analyse

Les données produites par l'annotation des éléments transposables sont générées avec les outils de la suite REPET et PanREPET (pangénomique).
Les données produites pour l'annotation de virus endogènes sont générées avec l'outil Caulifinder.

TYPE DE DONNEES

- Dataset
- Service
- Logiciel (décrit dans le mode de gestion dédié)

NATURE DES DONNEES

Les données d'entrée sont des séquences de génomes assemblées.
Les données de sortie sont des annotations d'éléments mobiles issus de prédictions automatiques et/ou de curations.

FORMAT DES DONNEES

- CSV/TSV
- GFF
- Fasta

PERIMETRE THEMATIQUE DES DONNEES

- Omics

DROITS DE PROPRIETE INTELLECTUELLE

QUI DETIENDRA LES DROITS SUR LES DONNEES ET LES AUTRES INFORMATIONS CREEES ?

- Les jeux de données générés par les membres de l'URGI sont la propriété d'INRAE. Cependant les membres de l'URGI sont responsables de la valorisation de leurs datasets.
- Les jeux de données générés par la plateforme dans le cadre de projets doivent suivre le PGD et les accords de consortium définis pour le projet.
- Les jeux de données générés dans le cadre de l'offre de service d'annotation proposée par la plateforme sont la propriété exclusive des demandeurs du service. Ils auront également le contrôle total sur leur valorisation.

Dans tous les cas, la plateforme doit être citée dans les remerciements de toute publication associée aux résultats obtenus de la manière suivante : *"This work was performed with the facilities of the Plant Bioinformatics Facility (https://doi.org/10.15454/1.5572414581735654E12)".* Les membres de l'URGI doivent également utiliser la double affiliation URGI/PlantBioinfoPF lors de la publication d'articles scientifiques. Ces prérequis sont détaillés sur notre site web, dans la page dédiée « How to cite ».

SENSIBILITE DES DONNEES

IDENTIFICATION DU NIVEAU DE SENSIBILITE DES JEUX DE DONNEES

- Diffusion limitée
  Sauf mention contraire, toutes les données gérées lors de l'analyse des génomes sont privées car elles sont soumises à des accords de consortium spécifiques, issues de partenaires privés et/ou sous embargo avant publication.
- Confidentiel
  Les données à caractère personnel soumises aux exigences de la CNIL et au RGPD sont confidentielles.
  Sur la plateforme, ces données concernent essentiellement des noms, prénoms et emails fournis lors de la souscription à une Offre De Service (ODS). Ces informations personnelles ne sont utilisées que par les membres de la plateforme pour traiter la demande, elles ne sont pas partagées avec des tiers. Elles sont conservées dans les meilleures conditions de sécurité et de confidentialité et archivées dans l'outil de gestion

de projet de l'URGI (JIRA) pendant la durée de vie des données scientifiques, tel que définie dans le présent PGD.
Conformément au règlement européen sur la protection des données personnelles (Règlement européen 2016/679), le propriétaire de données à caractère personnel dispose d'un droit d'accès, de rectification, d'opposition et de suppression des informations le concernant. La plateforme est assistée par le délégué à la protection des données personnelles (DPO) d'INRAE.

QUELLES SONT LES MESURES PRISES ET LES NORMES AUXQUELLES IL EST NECESSAIRE DE SE CONFORMER POUR GARANTIR LA SECURITE DES DONNEES SENSIBLES ?
Les analyses de génomes sont réalisées sur des environnements de recherche virtuel (ERV) via des machines virtuelles (VM) individuelles ou montées en cluster. Une clé SSH déclarée doit être fournie à l'équipe support de la plateforme afin d'accéder à la VM mise à disposition pour l'analyse et sur laquelle l'utilisateur se connecte avec un compte générique centos. En tant que propriétaire de cette VM il peut demander à en ouvrir l'accès à un autre utilisateur en passant par le support de la plateforme.
La VM est instanciée sur des serveurs INRAE hébergés dans le data center Ile-de-France.

S'IL Y A DES DONNEES A CARACTERE PERSONNEL, QUELLES SONT LES MESURES ENVISAGEES POUR LES PROTEGER AU COURS DU PROJET OU DANS LE CADRE D'UNE REUTILISATION ?
L'instance de JIRA utilisée par l'URGI pour gérer ses projets est installée sur des serveurs INRAE hébergés dans le data center Ile-de-France. L'accès aux données confidentielles qu'elle contient nécessite une authentification, les comptes étant créés à la demande par nos services internes.

PARTAGE DES DONNEES

Y A T'IL UNE OBLIGATION DE PARTAGE (OU A L'INVERSE UNE INTERDICTION OU UNE RESTRICTION) ?
Il n'y a aucune obligation de partage. La valorisation des données générées est sous la responsabilité du propriétaire de ces données, ou le cas échéant de la personne à l'origine de la demande. Il doit notamment être vigilants à respecter les lois sur les données de la recherche (open data, réglementation sur les données à caractère personnel).
Cependant les membres de l'URGI sont fortement encouragés et accompagnés par la plateforme à associer des DOI à leurs datasets et à les déposer en open access dans le dataverse URGI du portail Recherche Data Gouv.
La plateforme doit également être citée dans les remerciements de toute publication associée aux résultats obtenus de la manière suivante : *"This work was performed with the facilities of the Plant Bioinformatics Facility (https://doi.org/10.15454/1.5572414581735654E12)"*, comme précisé dans notre page dédiée « How to cite ».

QUELLES SONT LES REUTILISATIONS POTENTIELLES DE CES DONNEES ?
- Nouvelles analyses et/ou analyses plus poussées en relation avec les éléments transposables (exemple : impact des ET sur l'expression des gènes)
- A la fin de chaque prestation d'annotation d'éléments transposables, la plateforme demande au propriétaire si les jeux de données produits :
    - peuvent être réutilisés dans le cadre d'analyses interne à l'URGI
    - peuvent être mis à disposition de la communauté en open access via RepetDB (un SI dédiée aux éléments transposables) et/ou Recherche Data Gouv

LA LECTURE DES DONNEES NECESSITE-T-ELLE LE RECOURS A UN LOGICIEL OU UN OUTIL SPECIFIQUE ? SI OUI, LEQUEL ?
Non. Les données sont générées dans différents formats texte ouverts (CSV/TSV, GFF ...).

COMMENT LES DONNEES SERONT-ELLES PARTAGEES ?
Les données disponibles sur l'Environnement de Recherche Virtuel sont gérées par leur propriétaire. Si le demandeur du service souhaite partager ses données directement sur la VM avec quelqu'un d'autre, il doit demander au support de la plateforme d'ajouter la clé SSH du nouvel utilisateur sur cette VM. Sinon, les données peuvent être téléchargées et partagées en dehors de l'ERV sous la responsabilité du demandeur de service.

AVEC QUI ?
- Partenaire(s) identifié(s)

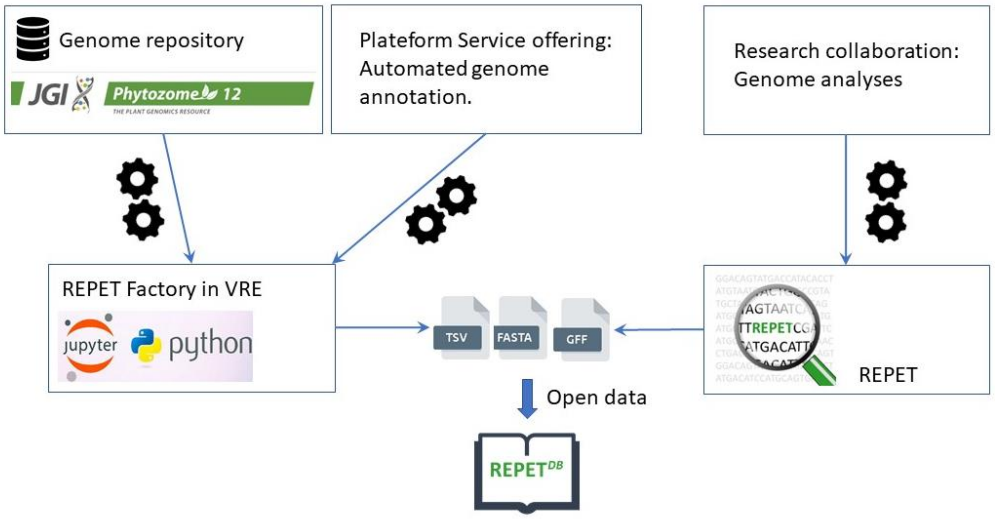
Les utilisateurs ont le contrôle total de leurs données, ils sont donc responsables du partage et de la valorisation de leurs datasets. Cependant, il y a une étape de validation de notre côté avant d'ajouter une nouvelle clé SSH sur une VM appartenant à un autre utilisateur.

SOUS QUELLE LICENCE ?
Les utilisateurs ont le contrôle total de leurs données, ils sont donc responsables du partage et de la valorisation de leurs datasets. Néanmoins, les données produites par la recherche publique ont vocation à être publiées sous licence ouvertes (LRPN 2016) et nous pouvons leur donner des conseils et les accompagner dans leurs démarches.

ORGANISATION ET DOCUMENTATION DES DONNEES

QUELS METHODES ET OUTILS SONT UTILISES POUR ACQUERIR ET TRAITER LES DONNEES, DEPUIS LEUR ACQUISITION JUSQU'A LEUR MISE A DISPOSITION, LEUR ARCHIVAGE OU LEUR DESTRUCTION ?
La plateforme fournit l'ERV sur son infrastructure OpenStack/Proxmox et elle en assure l'installation et la configuration. Le package REPET pour la détection et l'annotation des éléments transposables dans un génome, ainsi que ses dépendances, est installés par défaut sur ce serveur.
L'utilisateur dispose d'un accès administrateur sur son ERV, il est donc libre d'installer tout logiciel supplémentaire dont il a besoin, pour autant qu'il respecte notre charte d'utilisation. Cependant, la plateforme ne fournira de support que pour les logiciels qu'elle aura installés.

Les données sur les éléments transposables sont générées et gérées selon le schéma ci-dessous (https://urgi.versailles.inrae.fr/files/gnpis/data_integration_repet.jpg).



Les bibliothèques de consensus d'éléments transposables sont produites par REPET ou REPET factory dans des ERV. Les données ouvertes sont insérées dans RepetDB en utilisant des formats standards : fichier de classification (TSV), sortie Fasta et GFF de REPET. Les données sont ensuite rendues accessibles via RepetDB, un système d'information basé sur InterMine.

QUELLES METADONNE SERONT UTILISEES POUR ACCOMPAGNER LE JEU DE DONNEES ?
QUELS SERONT LES STANDARDS, VOCABULAIRES, TAXONOMIES... UTILISES POUR DECRIRE ET REPRESENTER LES DONNEES ET ELEMENTS DE METADONNEES ?
COMMENT LES METADONNEES SERONT-ELLES PRODUITES ET MISES A JOUR ?

Métadonnées	**Origine des métadonnées, mode de production**	**Standard et vocabulaire associés**	**Conditions et/ou fréquence de mise à jour**
Génomes assemblés	EBI Phytosome	Version de l'assemblage du génome au format fasta	Standard fixe (fasta) Mise à jour des données à chaque nouvelle version
Annotations de génome	Phytosome	Version de l'annotation du génome au format GFF3	Standard fixe (gff3) Mise à jour des données à chaque nouvelle version
Taxonomie	NCBI	Taxonomie de référence	Fixée par le NCBI

UNE DOCUMENTATION COMPLEMENTAIRE AUX METADONNEES EST-ELLE NECESSAIRE POUR DECRIRE LES DONNEES ET ASSURER LEUR REUTILISABILITE SUR LE LONG TERME ?
La suite REPET dispose de tutoriels expliquant comment utiliser l'outil et qui sont disponibles avec le code source (https://urgi.versailles.inrae.fr/Tools/REPET).
La documentation de Caulifinder est disponible sur une page dédiée du site web de la plateforme : https://urgi.versailles.inrae.fr/Tools/Caulifinder/CAULIFINDER.

COMMENT LES FICHIERS DE DONNEES SONT-ILS GERES ET ORGANISES : CONTROLE DES VERSIONS, CONVENTIONS DE NOMMAGE DES FICHIERS, ORGANISATION DES FICHIERS
Les données relatives à l'ERV sont gérées par leur propriétaire.

QUEL EST LE PROCESSUS DE CONTROLE QUALITE DES DONNEES ?
La suite REPET a ses propres étapes de contrôle de qualité (cf. https://urgi.versailles.inrae.fr/Tools/REPET).

STOCKAGE ET SECURITE DES DONNEES

LES SYSTEMES D'INFORMATION DE LA STRUCTURE ONT-ILS FAIT L'OBJET D'UNE ANALYSE DE RISQUES OU D'UNE HOMOLOGATION ?
Nos SI ne sont pas encore homologués, mais une démarche d'analyse de risques est en cours dans le cadre de la mise en place de la politique de sécurité des SI d'INRAE.

QUELS TYPES DE SUPPORTS PHYSIQUES SONT UTILISES POUR STOCKER LES DONNEES ?
Les données, gérées sur l'ERV (données copiées et/ou générées) sont stockées sur des volumes CEPH montés sur la VM. Les données générées à toutes les étapes de REPET sont gérées dans une base de données MySQL dédiée qui est hébergée sur notre infrastructure OpenStack/Proxmox. Le tout est hébergé dans le data center Ile-de-France d'INRAE.

QUELLES SONT LES MESURES DE SECURITE MISES EN PLACE LORS DES ETAPES DE TRANSFERT DES DONNEES ?
Les données pertinentes sont copiées dans l'ERV à l'aide d'outils cryptographiques, tels que des sessions shell sécurisées (SSH).

QUELLE EST LA VOLUMETRIE ACTUELLE ET PREVISIONNELLE ?
Le volume total disponible est d'environ 70 TB en juillet 2021.
Il peut être augmenté facilement si besoin (l'infrastructure est facilement adaptable facilement et les ressources nécessaires sont disponibles).

L'ENTITE HEBERGEANT PHYSIQUEMENT LES DONNEES A-T-ELLE UNE POLITIQUE DE SECURITE DE L'INFORMATION ET A-T-ELLE UN PLAN D'ASSURANCE SECURITE ?
Notre politique de sécurité suit la charte des infrastructures de recherche d'INRAE.
Les données gérées sur l'ERV sont stockées sur les serveurs de l'URGI, hébergés dans le data center Ile-de-France avec un contrôle d'accès physique strict. Les volumes CEPH et la VM OpenStack/Proxmox ne sont pas sauvegardés à ce stade, mais les données existent en trois copies sur CEPH et la reconstruction des VM nues pour l'ERV est automatisée.

SECURITE - CONFIDENTIALITE : LES DONNEES FONT-ELLES L'OBJET D'ECHANGE OU DE PARTAGE AVEC DE TIERS ACTEURS ET SELON QUELLES MODALITES ? COMMENT SONT DETERMINES LES DROITS D'ACCES AUX DONNEES AVANT LEUR PUBLICATION ?
Une clé SSH déclarée doit être fournie pour accéder à l'ERV mis à disposition par la plateforme. Seul le souscripteur du service peut demander à ouvrir l'accès à son ERV et à ses données à un autre utilisateur.
Les données peuvent être accédées à l'aide d'outils cryptographiques, tels que des sessions shell sécurisées (SSH) pour copier des fichiers. L'accès à ces fichiers est contrôlé à l'aide de permissions UNIX bien connues sur des serveurs Linux, les volumes de données CEPH sont uniquement accessibles sur la VM des utilisateurs et ne peuvent pas être montés ailleurs.

SECURITE - INTEGRITE - TRACABILITE : QUELLES SONT LES MESURES DE PROTECTION MISES EN ŒUVRE POUR SUIVRE LA PRODUCTION ET L'ANALYSE DES DONNEES ?
Les VM sur l'OpenStack/Promox de l'URGI sont soutenues par la technologie CEPH et bénéficie de ses fonctionnalités d'intégrité et de sécurité propres telles que le scrubbling (triple réplication des données sur différents stockages physiques, comparaison régulière des données répliquées) ou les vérifications des cycles de redondances (CRC recalculant l'empreinte des données à leur lecture et en la comparant à celle stockée lors de l'écriture).

LES AGENTS DE LA STRUCTURE ONT-ILS BENEFICIE D'UNE SENSIBILISATION AUX BONNES PRATIQUES D'HYGIENE NUMERIQUE ?
Oui.
Tous les agents ont reçu les communications de la cellule SSI INRAE. Certains d'entre eux, en première ligne, ont également suivi une formation à la sécurité informatique offensive et défensive et sont en lien avec la chaîne SSI INRAE.

ARCHIVAGE ET CONSERVATION DES DONNEES

QUELLES SONT LES DONNEES A CONSERVER SUR LE MOYEN OU LE LONG TERME ET QUELLES SONT LES DONNEES A DETRUIRE ?
À la fin de la période d'allocation des ressources, l'ERV est fermé et les données sont effacées dans un délai de 15 jours. Comme il n'y a pas d'archivage, les utilisateurs doivent récupérer les données qu'ils souhaitent conserver avant cette échéance.
Pour les membres de l'URGI, l'ERV et ses données sont conservés aussi longtemps que nécessaire sur un espace de stockage monté et sauvegardé.

SUR QUELLE PLATEFORME D'ARCHIVAGE PERENNE SERONT ARCHIVEES LES DONNEES A CONSERVER SUR LE LONG TERME ?
SINON, QUELLES PROCEDURES SERONT MISES EN PLACE POUR LA CONSERVATION A LONG TERME ?
Il n'y a pas d'action d'archivage des données. Cependant, nous proposons une publication des jeux de données dans RepetDB (un système d'information dédiée aux éléments transposables) et/ou le dataverse URGI du portail Recherche Data Gouv, ce dernier garantissant un accès aux données pendant 10 ans grâce à l'attribution d'un DOI.

QUELLE EST LA DUREE DE CONSERVATION DES DONNEES ?
À la fin de la période d'allocation des ressources, les utilisateurs disposent de 15 jours pour récupérer les données qu'ils souhaitent conserver.

QUELLES GARANTIES DE FINANCEMENTS COUVRIRONT LES COUTS ASSOCIES A LA CONSERVATION A LONG TERME ?
INRAE fournit des financements de base (ressources humaines et soutien financier de la plateforme) via les départements de recherche BAP et ECODIV ainsi que la CNOC (Commission Nationale des Outils Collectifs).
La plateforme a également accès à des financements spécifiques (ODS payantes).

GLOSSAIRE

ANR	Agence Nationale de la Recherche
API (RESTful)	Application Programing Interface (REpresentational State Transfer)
BAM	Binary Alignment Map (format de fichier)
BAP	Département INRAE Biologie et Amélioration des Plantes (https://www.inrae.fr/departements/bap)
BED	Browser Extensible Data (format de fichier)
BrAPI	Breeding API (https://brapi.org/)
CNIL	Commission Nationale de l'Informatique et des Libertés (https://www.cnil.fr/)
CSV	Coma Separated Values (format de fichier)
DOI	Digital Object Identifier
DPO	Data Protection Officer
ECODIV	Département INRAE Écologie et Biodiversité (https://www.inrae.fr/departements/ecodiv)
ERV/VRE	Environnement de Recherche Virtuel / Virtual Research Environment
ET/TE	Eléments Transposables / Transposable Elements
ETL	Extract Transform Load
FP7	7th Framework Programme for Research
GFF	General feature format (format de fichier)
GWAS	Genome Wide Association Study
HE	Horizon Europe
HTML	Hypertext Markup Language
HTTP(S)	Hypertext Transfer Protocol (Secure)
IBISA	Infrastructures en Biologie Sante et Agronomie (https://www.ibisa.net/)
IFB	Institut Français de Bioinformatique (https://www.france-bioinformatique.fr/)
INRAE	Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement (https://www.inrae.fr/)
ISA-Tab	Investigation/Study/Assay Tab-delimited
ISC	Infrastructure Scientifique Collective
IWGSC	International Wheat Genome Sequencing Consortium (https://www.wheatgenome.org/)
JSON	JavaScript Object Notation (format de fichier)
LDAP	Lightweight Directory Access Protocol
MCPD	Multi-Crop Passport Descriptors (https://www.fao.org/plant-treaty/tools/toolbox-for-sustainable-use/details/en/c/1367915/)
MIAPPE	Minimal Information About plant Phenotyping Experiments (www.miappe.org)
PGD	Plan de Gestion de Données
PIA	Plan d'Investissement d'Avenir
QTL	Quantitative Trait Loci
RDA	Research Data Alliance
RDF	Resource Description Framework (format de fichier)
RGPD	Règlement Général sur la Protection des Données (https://www.cnil.fr/fr/rgpd-de-quoi-parle-t-on)
SAM	Sequence Alignment Map
SCSS	Syntactically Awesome Style Sheet (format de fichier)
SGBD	Système de Gestion de Base de Données
SGBDR	Système de Gestion de Base de Données Relationnel
SI	Système d'Information
SQL	Structured Query Language (language)
SSH	Secure Socket Shell
TSV	Tabulated Separated Values (format de fichier)
URGI	Unité de Recherche en Génomique-Info (https://www6.versailles-grignon.inrae.fr/urgi/)
VCF	Variant Call Format (format de fichier)
VM	Virtual Machine
YAML	Yet Another Markup Language (format de fichier)