URGI Plant Bioinformatics Facility (PlantBioinfoPF) data management plan

Célia Michotey, Michael Alaux, Raphaël Flores, Anne-Françoise Adam-Blondon

URGI PLANT BIOINFORMATICS FACILITY (PLANTBIOINFOPF) DMP

Data management plan created using DMP OPIDoR, based on the "INRAE – Modèle Structure" model provided by INRAE – French National Research Institute for Agriculture, Food and Environment.

Table of contents

URGI Plant Bioinformatics facility (PlantBioinfoPF) DMP 1
Management plan informations 2
Structure informations 3
Management mode « Information Systems » 4
Data general presentation 4
Intellectual property rights 5
Data sensitivity 5
Data sharing 6
Data organisation and documentation 7
Data storage and security 9
Archiving and data retention 10
Management mode « Software » 12
Data general presentation 12
Intellectual property rights 13
Data sensitivity 13
Data sharing 13
Data organisation and documentation 14
Data storage and security 15
Archiving and data retention 15
Management mode « Genome analyses » 17
Data general presentation 17
Intellectual property rights 17
Data sensitivity 17
Data sharing 18
Data organisation and documentation 19
Data storage and security 20
Archiving and data retention 20
Glossary 22

MANAGEMENT PLAN INFORMATIONS

PLAN
Title: PlantBioinfoPF DMP
Language: English
Creation date: 16/02/2021
Last modification date: 04/06/2024

PLAN IDENTIFIER
https://doi.org/10.15454/9HM5UI

PLAN LICENSE

SUMMARY
The URGI, an INRAE research, hosts a bioinformatics platform (Plant Bioinformatics Facility - doi:10.15454/1.5572414581735654E12) that supports research activities in plant genetic and genomic. The platform's services cover database design, software engineering, software hosting, data integration and training. The platform's activities benefit from URGI's research activities (data integration, repeat annotation, study of genome structure and evolution). It belongs to IFB ("Institut Français de Bioinformatique"), the french node of the Elixir European network of bioinformatics platforms. It has been certified by INRAE's ISC ("Infrastructure Scientifique Collective") committee and by the GIS IBISA ("Groupement d'Intérêt Scientifique - Infrastrutures en Biologie Sante et Agronomie") as a national strategic platform. The platform also belongs to the INRAE Research Infrastructure BioinfOmics and is part of the Saclay Plant Science network and the Biosphera Graduate School. It is ISO9001 v. 2015 certified.
For more information, visit our website: https://urgi.versailles.inrae.fr.

Research fields (according to OECD classification): Biological sciences (Natural sciences), Computer and information sciences
Funding sources: INRAE - National Research Institute for Agriculture, Food and the Environment
Partners :

- INRAE department BAP: "Biologie et Amélioration des Plantes"
- INRAE department ECODIV: "Ecologie et Biodiversité"

STRUCTURE MANAGEMENT METHODS

Management mode	Description	Type
Information Systems (IS)	DMP of the data managed in our Information Systems (GnpIS and data portals)	Dataset
Software	DMP of the source code of the software we develop	Software
Genome analysis	DMP of the data managed as part of our genome analysis	Dataset

CONTRIBUTORS

Last name, first name	ORCID	Role
Alaux, Michael	https://orcid.org/0000-0001-9356-4072	Project coordinator
Flores, Raphael	https://orcid.org/0000-0002-0278-5441	Project coordinator, Infrastructure contact
Michotey, Célia	https://orcid.org/0000-0003-1877-1703	DMP manager Contact for IS datasets and software
Pommier, Cyril	https://orcid.org/0000-0002-9040-8733	Contact for IS datasets and software
Confais, Johann	https://orcid.org/0000-0003-2945-5036	Contact for genome analysis datasets and software

VERSIONS HISTORY

Date	Version	Status	Author	Validated by
16/02/2021	1	Published	Célia Michotey	Anne-Françoise Adam-Blondon
04/06/2024	2	Published	Célia Michotey	Michael Alaux

STRUCTURE INFORMATIONS

STRUCTURE NAME
Plant Bioinformatics Facility (PlantBioinfoPF)

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

PlantBioinfoPF is part of BioinfOmics research infrastructure

ANSci card (restricted access): https://actif-numerique.inrae.fr/ansci/app/systeme-information/449

STRUCTURE IDENTIFIER
https://doi.org/10.15454/1.5572414581735654E12

STRUCTURE RESPONSABILITIES

Last name, first name	Email	Role
Alaux, Michael	michael.alaux@inrae.fr	Scientific manager Deputy unit manager
Flores, Raphael	raphael.flores@inrae.fr	Operational manager
Adam-Blondon, Anne-Françoise	anne-francoise.adam-blondon@inrae.fr	Unit manager

GUARDIANSHIP INSTITUTION(S)
INRAE - French National Research Institute for Agriculture, Food and Environment

INRAE DEPARTEMENT OF AFFILIATION
BAP: Biologie et amélioration des plantes
The platform also includes ECODIV: Ecologie et Biodiversité staff (1.8 FTE).

FUNDER(S) (ENABLING DATASETS ACQUISITION – OUTSIDE PROJECTS)
Datasets are currently acquired as part of collaborative projects with platform members. They benefit from various types of funding, the main ones being: INRAE; ANR (PIA or not); EU (FP7, H2020, HE).
Our national or international data providers (e.g. IWGSC) outside collaborative projects do not provide us with their funding sources.

MANAGEMENT MODE « INFORMATION SYSTEMS »

The information systems made available by PlantBioinfoPF are:

- The GnpIS data warehouse

GnpIS is an integrative, multi-species information system (IS) dedicated to plants and their pests. It enables researchers to access and cross-reference genetic data (accessions, phenotypes, markers, QTLs, polymorphisms, association genetics) and genomic data (sequences, physical maps, genome annotations and expression data) for species of agronomic and forestry interest. The IS is accessible via a web portal, and enables different types of data to be browsed, either independently via dedicated interfaces, or simultaneously using search tools.

- Federations of data portals

These research portals, based on the same tool, facilitate discovery and access to FAIR (Findable, Accessible, Interoperable, Reusable) data through a federation of distributed IS on a national (RARe and BRC4Env portals, dedicated to the BRC community), European (FAIDARE, dedicated to the ELIXIR Plant Sciences community) and international (WheatIS Data Discovery, dedicated to the wheat community) scale. They aim to provide researchers with quick and easy access to relevant biological data using specific keywords and filters.

The management of the codes of the tools linked to these IS is described in the dedicated management mode: "Software".

DATA GENERAL PRESENTATION

DATA OBTENTION MODE

- Data produced by a third party
- Data generated by the structure:
  - Genome annotations (genes and transposable elements from automatic prediction and/or curations)
  - Trait ontologies

ORIGIN

- Analysis
- Aggregation
- Experimentation
- Observation

GnpIS stores and integrates genomic data, genetic data and phenotypic data from INRAE researchers and their national and international partners.

DATA TYPE

- Collection
- Dataset
- Software (described in dedicated management mode)

NATURE OF THE DATA

Plant genomic and genetic data are provided by INRAE units, their projects partners and the International Wheat Genome Sequencing Consortium (IWGSC).
GnpIS stores mainly textual data:

- Genetic resources (accessions, passport data, images)
- Genomic data (mainly genome annotations, polymorphisms and synteny)
- Genetic data (QTLs and GWAS analysis)
- Phenotypic data (ontologies, observations and experimental data)

DATA FORMAT

Collections (genetic resources)

- CSV/TSV
- XLS/XLSX
- JSON

Datasets

- CSV/TSV
- XLS/XLSX
- VCF

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

GnpIS has begun to expose some of its data in a semantic representation to improve data integration with other databases (RDF format, see https://urgi.versailles.inra.fr/About-us/News/2017/RDF-Phenotyping).

DATA THEMATIC PERIMETER

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

INTELLECTUAL PROPERTY RIGHTS

WHO WILL OWN THE RIGHTS ON THE DATA AND OTHER INFORMATION CREATED?

Our policy is described in our terms of use. We also describe the conditions specific to certain datasets on dedicated web pages.

URGI encourages its users to associate DOIs with deposited datasets. It helps data producers to obtain these DOIs and associate them with the appropriate metadata, in collaboration with the French open data portal Recherche Data Gouv. This enables us to discuss the license to be associated with the data, with a proposal for CC-BY V4.0 by default for public data.

More generally, we implement in GnpIS the terms described in the data management plans of the consortium agreements of the projects producing the data to be stored, some data being accessible only to defined consortia of users.

The decision to publish data is taken with the depositor.

DATA SENSITIVITY

IDENTIFYING THE DATASETS SENSITIVITY LEVEL

- Public data
  The majority of data integrated into GnpIS is public data made available as part of open science. Data available via federated portals are all public.
- Limited distribution
  Some of the data integrated into GnpIS are private, as they are subject to specific consortium agreements, originate from private partners and/or are embargoed before publication.
- Confidential
  Personal data subject to CNIL and GDPR requirements are confidential. On the platform, this data mainly concerns surnames, first names and emails.
  - Personal information provided when requesting a service is only used by platform members to process this request, it is not shared with third parties. This information is kept in the best conditions of security and confidentiality and archived in URGI's project management tool (JIRA) for the life of the scientific data, as defined in this DMP.
  - Personal information provided when ordering genetic resources (via FAIDARE, RARe and BRC4Env order basket) is used only by the BRC managers who will process this request. It is not shared with third parties. This information is kept in the best conditions of security and confidentiality in a dedicated database, as defined in the present DMP.
  - Personal data included in datasets enable information traceability. They are therefore published under the same status (public or private) as the associated scientific data when they are integrated into GnpIS. They will be managed in the same way, as described in this DMP.

  In accordance with the European regulation on the protection of personal data (European Regulation 2016/679), the owner of personal data has the right to access, rectify, oppose and delete information concerning him or her. The platform is assisted by INRAE's personal data protection delegate (DPO).

WHAT MEASURES ARE TAKEN AND WHAT STANDARDS MUST BE MET TO GUARANTEE THE SECURITY OF SENSITIVE DATA?

The JIRA instance used by URGI to manage its projects, our IS instances (GnpIS, FAIDARE, RARe and BRC4Env) and the databases on which they are based are all installed on INRAE servers hosted in the Ile-de-France data center. Access to these tools and the sensitive data they contain requires authentication via the Apache HTTP authentication system, with accounts created on demand by our internal services.

In the case of GnpIS, a specific group including all persons who can access a private dataset (e.g. colleagues, project partners) is defined with the owner of the data in the IS. Data is tagged in the database to this group with defined access rights, so that only people belonging to the group can access it. Authentication on the web interface is governed by Apache, following the authorizations defined in the database and described above.

IF PERSONAL DATA IS INVOLVED, WHAT MEASURES ARE PLANNED TO PROTECT IT DURING THE PROJECT OR IN THE EVENT OF REUSE?

Confidential data is managed in the same way as sensitive data: instances of the tools used are installed on INRAE servers hosted in the Ile-de-France data center, and access to the tools and the data they contain requires authentication controlled by our internal departments.

The transfer of confidential data to third parties is subject to validation and follow-up by the data owner. It follows the same procedure as described above, reusing the authentication and authorization systems under our control.

DATA SHARING

IS THERE AN OBLIGATION TO SHARE (OR A PROHIBITION OR RESTRICTION)?

The data provider is committed to publicly open the data. An embargo period may be defined for use by the scientific community.

The decision to publish data is taken with the scientific manager at the time of submission.

However, the platform should be cited in any publication citing the integrated dataset(s): *"This work was performed with the facilities of the Plant Bioinformatics Facility (https://doi.org/10.15454/1.5572414581735654E12)".*

WHAT ARE THE POTENTIAL REUSES OF THIS DATA?

- Update of international genetic resource catalogs
- Update of other information systems
- Reuse for new research purposes
- Support for public policy, expertise
- Trainings

DOES DATA READING REQUIRE THE USE OF SPECIFIC SOFTWARE OR TOOLS?
IF SO, WHICH ONE?

Data can be accessed via IS web interfaces (GnpIS, FAIDARE, WheatIS search, RARe & BRC4Env) and via standardized web services (RESTful APIs) that enable automatic access via programming.

Data can be downloaded in various standard text formats (CSV/TSV, GFF, VCF, JSON …).

HOW DATA WILL BE SHARED?

Data are shared via web interfaces and web services APIs provided by IS. Dara can be downloaded in various standard text formats (CSV/TSV, GFF, VCF, JSON …).

Authentication is required to access private data (see our terms of use).

Some datasets are also available in the URGI dataverse on the French open data portal Recherche Data Gouv.

WITH WHO?

- Public data are freely accessible, so it can be shared with anyone (open access)
- Private data are shared with identified users (academic and/or private partners)

UNDER WHICH LICENCE?

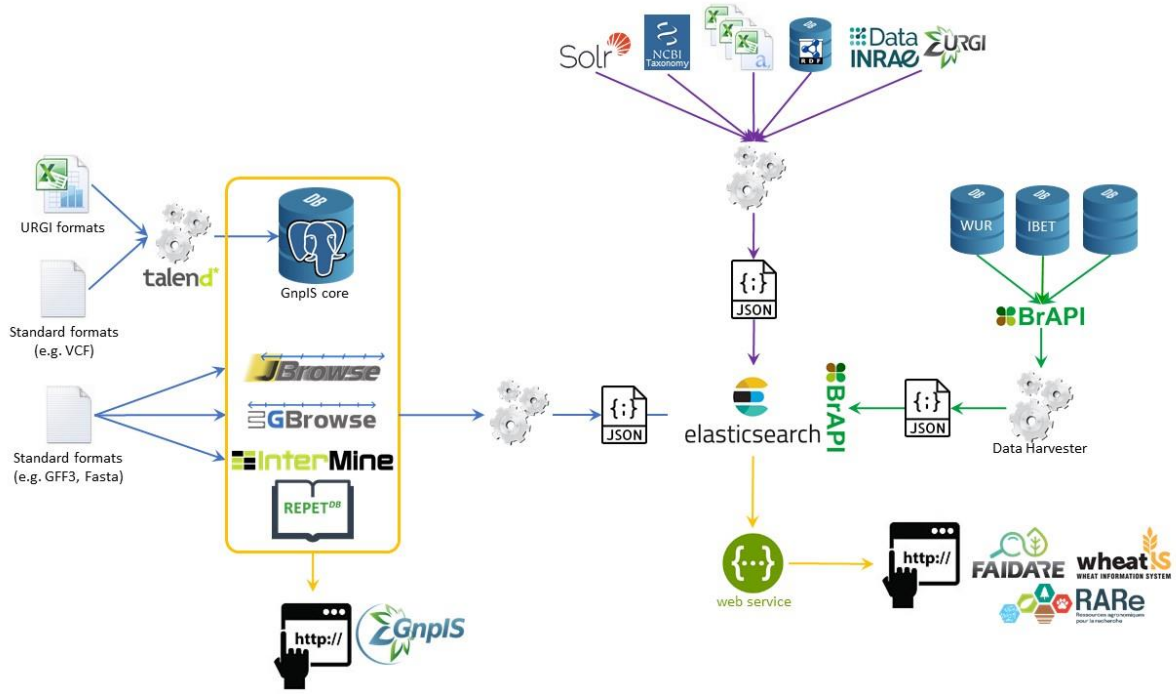
We encourage and support scientific managers to associate a DOI with their datasets via a publication in the URGI dataverse on the French open data portal Recherche Data Gouv. This enables us to discuss the license to be associated with the data.

By default, public data are licensed under CC-BY V4.0.

DATA ORGANISATION AND DOCUMENTATION

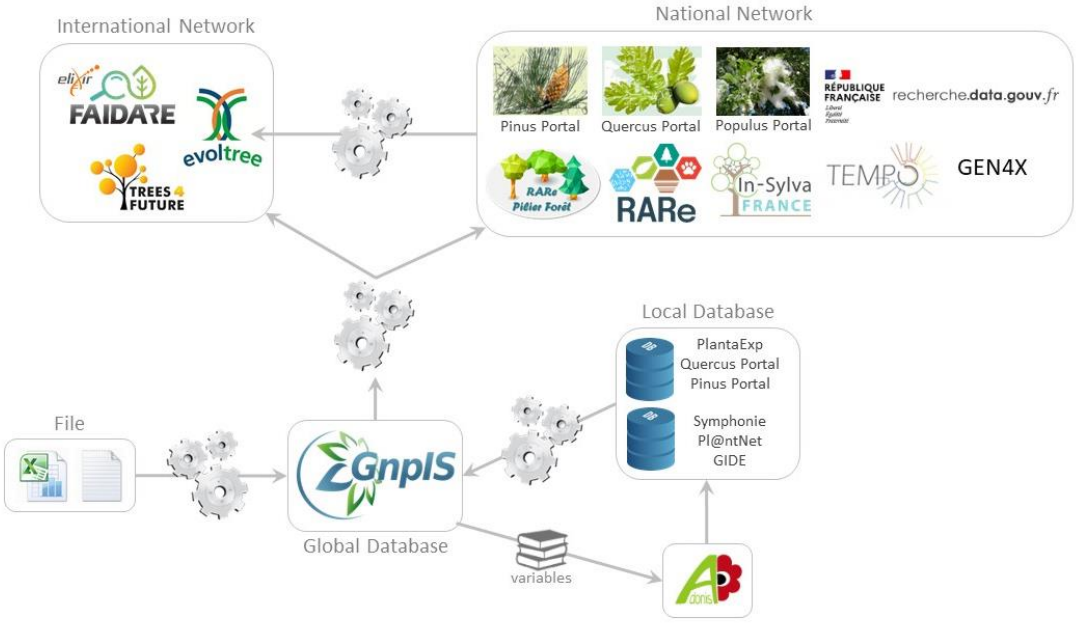
WHAT METHODS AND TOOLS ARE USED TO ACQUIRE AND PROCESS DATA, FROM ACQUISITION TO AVAIBILITY, ARCHIVING OR DESTRUCTION?

Quality procedures (ISO9001 v. 2015, restricted access) explain in detail how to manage data in the platform's various IS (acquisition, validation, integration and visualization in the IS, archiving).
The diagram below illustrates data management (https://urgi.versailles.inrae.fr/files/gnpis/data_integration_gnpis.jpg).



Data produced by research teams are submitted to PlantBioinfoPF using standard and in-house exchange formats. ETL – Extract Transform Load tools (mainly Talend Open Studio projects and Perl and Python scripts) are used to check data quality and consistency and to populate GnpIS databases (PostgreSQL and MySQL). Data from GnpIS and other IS to be displayed in the federated portals are integrated together in an Elasticsearch database (document-oriented NoSQL) using ETL tools (mainly Bash and Python scripts). The data are then made accessible via RESTful APIs and web interfaces (GnpIS, FAIDARE, WheatIS search, RARe & BRC4Env, Siregal).

In the case of forest trees, different types of data are managed in different IS with different accessibility. To share this knowledge, we set up an automated data flow to synchronize data shared by several IS, as shown in the schema below (https://urgi.versailles.inrae.fr/files/gnpis/forest_interoperability.jpg).



Data produced by research teams are managed in local IS: data acquisition tools, such as Adonis, local databases and files (standards or in-house formats). When data are ready to be shared, they are extracted from these local sources and inserted into GnpIS, our global IS. When relevant, data can then be extracted from GnpIS for insertion into other IS to enhance their visibility on a national and/or international scale. ETL tools specific to each IS are used to automatically manage these data flows.

WHAT METADATA WILL BE USED TO ACCOMPANY THE DATASET?
WHAT STANDARDS, VOCABULARIES, TAXONOMIES... WILL BE USED TO DESCRIBE AND REPRESENT DATA AND METADATA?
HOW WILL METADATA BE PRODUCED AND UPDATED?

Metadata	**Metadata origin and production mode** *(e.g. manual input, automatic annotation)*	**Standard and associated vocabularies**	**Conditions or frequency update (if applicable)** *(e.g. change in accessibility)*
Genetic resources passport data	GnpIS submission format filling (semi-manual) by resource managers	• Multi-Crop Passport Descriptors (MCPD) • Referential taxonomies (NCBI, TAXREF, Catalogue of Life …) • Trait ontologies (CropOntology Trait dictionary, Trait Ontology …)	• MCPD is governed by the FAO • Trait dictionaries updates depends on the CropOntology communities
Plant phenotyping experiments	GnpIS submission format filling (semi-manual) by data producers	• Minimal Informations About Plant Phenotyping Experiments (MIAPPE) • Breeding API (BrAPI) • CropOntology Trait dictionary	• www.miappe.org • https://brapi.org/ • Trait dictionary updates depends on the communities
Genomic data	GnpIS submission format filling (semi-manual) by data producers	EMBL, NCBI recommendations	

IS ADDITIONAL DOCUMENTATION REQUIRED TO DESCRIBE DATA AND ENSURE LONG TERM REUSABILITY?
GnpIS submission formats ensure that the metadata essential for describing and reusing dataset is available. If necessary, documents can be attached to the data to provide further information (e.g. readme files), non-proprietary

formats being strongly recommended. Moreover, we are increasingly associating DOI with datasets and we recommend the publications of data papers to our users to enhance this documentation.
Data can also be retrieved in standard formats:

- Trait ontologies can be retrieve in various formats following the CropOntology Trait dictionary standard (Excel, CSV, JSON) via GnpIS web interface or via a web service BrAPI compliant (Breeding API, a standardized RESTful API that enables interoperability between plant breeding databases).
- Accessions and their passport data can be retrieved in a CSV format compliant with MCPD (Multi-Crop Passport Descriptors, a standard that facilitates the exchange of germplasm information) via GnpIS web interface or a BrAPI compliant web service.
- Phenotyping data can be retrieved as a MIAPPE (Minimal Informations About Plant Phenotyping Experiments, a standard specifying the metadata needed to properly describe plant phenotyping experiments) compliant ISA-Tab format or computationally via a BrAPI compliant web service.
- Polymorphism data can be retrieved as a VCF and CSV formats via GnpIS web interface.
- Genome annotation data can be retrieved as a GFF3 format via species dedicated web pages when GnpIS is the official repository, or from international archives as described in the metadata of GnpIS JBrowse instances.
- Genetic data can be retrieved in a CSV format, as there is no internationally agreed standard.

Tools are available to produce these metadata:

- List of Digital Curation Center metadata tools
- RDA (Research Data Alliance) metadata directory

HOW DATA FILES ARE MANAGED AND ORGANISED: VERSION CONTROL, FILE NAMING CONVENTIONS, FILE ORGANISATION

Quality procedures (restricted access) exist to define in details how to manage data in our IS (GnpIS and federated portals).
ETL tools (Bash, Perl or Python scripts and Talend Open Studio projects) are used to process data files and control data quality and integrity throughout its management, from submission to access in GnpIS portal. Implementation of reproducible ETL workflows is currently under development.
The trait ontologies we develop and maintain are managed in version management software, more specifically on a public forgemia project which is the GitLab forge of INRAE's MATHNUM department (see Ontologies).

WHAT IS THE DATA QUALITY CONTROL PROCESS?

Quality procedures (restricted access) exist to define in details how to manage data in our IS (GnpIS and federated portals), including the steps to control data quality and integrity.
Steps to check the validity of the submitted files (e.g. format and mandatory content respected), data quality (e.g. controlled vocabularies/ontologies correctly used) and dataset consistency (e.g. consistency of data and links between them) are implemented in the ETL tools (Bash, Perl or Python scripts and Talend Open Studio projects) we use to manage datasets in our IS. Moreover, the use of relational databases for GnpIS (PostgreSQL and MySQL RDBMS) guarantees the integrity, based on some business constraints that have been put in place over the years. Finally, when an insertion is made, several checks are also carried out on the number of entries to ensure that no data is missing or duplicated.

DATA STORAGE AND SECURITY

HAVE THE STRUCTURE'S INFORMATION SYSTEMS BEEN SUBJECTED TO RISK ANALYSIS OR CERTIFICATION?

Our information systems are not yet certified, but a risk analysis process is underway as part of the implementation of INRAE's IS security policy.

WHAT TYPES OF PHYSICAL MEDIA ARE USED TO STORE DATA?

Data producers can access the data repository service in various ways, which are described on the URGI website:

- Via the platform service-offering
- Via GnpIS submission form

And specific pages dedicated to the wheat community:

- Wheat Initiative's WheatIS expert group
- IWGSC data repository

Submitted files are stored on a NetApp/CEPH volume mounted on URGI servers. The content of the files is processed so that the data can be inserted into the databases making up our information systems (MySQL and PostgreSQL RDBMS, NoSQL Elasticsearch). These databases are hosted on virtual machines (VMs) in our virtualization instances (Proxmox and OpenStack cloud). Everything is hosted in INRAE's Ile-de-France data center.

The trait ontologies we develop and maintain are versioned on a public project of the GIT forge forgemia (see Ontologies).

WHAT SECURITY MEASURES ARE IN PLACE FOR DATA TRANSFER?
Data downloaded from our IS (web interfaces and services) are transferred over the network using the encrypted HTTPS protocol. If the data are private, authentication governed by the Apache HTTP authentication system and under our control is required before downloading can begin.
Data versioned in the forgemia GIT can be retrieved via the secure HTTPS or SSH protocols.

WHAT IS THE CURRENT AND PROJECTED DATA VOLUME?
Datasets: 6.5 TB of storage as of June 2024
Trait ontologies: 116 MB in June 2024

DOES THE ENTITY PHYSICALLY HOSTING THE DATA HAVE AN INFORMATION SECURITY POLICY AND A SECURITY ASSURANCE PLAN?
Our security policy follows the INRAE's Research Infrastructures Charter.
All submitted data and attached files are stored on URGI servers, hosted in INRAE's Ile-de-France data center, and regularly backed up. The databases making up our IS are backed up via dumps twice a month with a 2-month retention period. NetApp storage volumes are backed up by snapshot twice a day. These backups are duplicated monthly at the INRAE data center in Toulouse.

SECURITY - CONFIDENTIALITY: IS DATA EXCHANGED OR SHARED WITH THIRD PARTIES AND HOW? HOW ARE DATA ACCESS RIGHTS DETERMINED BEFORE PUBLICATION?
Private data can only be accessed after an authentication process. Access rights are given to a specific group comprising all the people who can access the data (e.g. colleagues, project partners) and defined with the data owner. Data is tagged in GnpIS with this group, so that only people belonging to this group can access it.
The transfer of private data to third parties is subject to validation and follow up by the data owner.

SECURITY - INTEGRITY - TRACABILITY: WHAT SAFEGUARDS ARE IN PLACE TO MONITOR DATA PRODUCTION AND ANALYSIS?
Private data can be provided using cryptographic based tools, such as secure shell sessions to copy data to our servers, which are supported by NetApp/CEPH technology, which is configured to take a snapshot of all copied data twice a day. Access to copied files is controlled using UNIX permissions on Linux servers, which mount NetApp/CEPH volumes only where necessary.
Submitted files follow the quality procedures (restricted access) already mentioned, and all steps are recorded in our internal instance of JIRA (digital lab notebook) in a dedicated task.
Trait Ontologies are managed using the functionalities of a GIT forge (forgemia).

HAVE THE STRUCTURE'S AGENTS BEEN MADE AWARE OF GOOD DIGITAL HYGIENE PRACTICES?
Yes.
All staff have received communications from the INRAE SSI unit. Some front-line staff have also received training in offensive and defensive computer security, and are in contact with the INRAE SSI chain.

ARCHIVING AND DATA RETENTION

WHICH DATA SHOULD BE KEPT FOR THE MEDIUM OR LONG TERM AND WHICH SHOULD BE DESTROYED?
Data to be preserved over the long term are:

- Genetic resources passport data and associated images
- Processed phenotyping data (not raw sensor data files)
- The Trait Ontologies we manage
- Annotation data for transposable elements

ON WHICH PERMANENT ARCHIVING PLATFORM DATA WILL BE STORED FOR LONG TERM RETENTION? WHAT PROCEDURES WILL BE IMPLEMENTED FOR LONG TERM PRESERVATION?
Trait Ontologies are versioned on a public repository in the forgemia GitLab (see Ontologies). This repository has been synchronized with Software Heritage (an initiative to preserve and share public source code) to enable the data it contains to be automatically archived at regular intervals (see software heritage Ontologies).
Otherwise, there is no archiving action beyond GnpIS itself. However, we are gradually imposing parallel publication of datasets in the URGI dataverse of the Recherche Data Gouv portal, which, in conjunction with the DOI, guarantees access to the data for 10 years.

HOW LONG WILL THE DATA BE KEPT?

Trait Ontologies will be publicly accessible for as long as forgemia (or an equivalent) exists. An alternative solution will be used in the unlikely event that these solutions are shut down.

Genetic resources and phenotyping data are strategic and will be conserved as long as INRAE can provides the means to do so.

TE annotations are also versioned and kept for as long as possible, as there is no central open-access repository.

For other data, this depends on the interest in being integrated into GnpIS with other datasets, and on strategic issues to be discussed with the data provider.

WHAT FUNDING GUARANTEES WILL COVER THE COSTS ASSOCIATED WITH LONG-TERM PRESERVATION?

The platform's IS are strategic assets of INRAE, which provides core funding (human resources and financial support for the platform) via the BAP and ECODIV research departments and the CNOC (“Commission Nationale des Outils Collectifs”).

The platform also has access to project funding to support these IS.

MANAGEMENT MODE « SOFTWARE »

The software developed and made available by PlantBioinfoPF are:

- Tools linked to our information systems (IS)
  - GnpIS data warehouse and FAIDARE, WheatIS Data Discovery, RARe and BRC4Env data federation portals
  - ETL (Extract Transform Load) tools used to integrate data into our IS
  - RARe-Basket (restricted access), a management tool for ordering genetic resources
  - Trait ontology widget, a tool for visualizing ontologies compatible with the CropOntology standard.
- Genome analysis tools, such as the REPET suite and TE_finder for annotating transposable elements, Caulifinder for annotating endogenous viral elements (EVE) of caulimoviride origin.

Data management in these tools are described in the dedicated management modes: “Information systems” and “Genome analysis”.

DATA GENERAL PRESENTATION

DATA OBTENTION MODE

- Data produced by a third party
- Data generated by the structure

As the code of our software is open, third parties can contribute and add new functionalities.

ORIGIN

- Code

DATA TYPE

- Software
- Workflow
- Service

NATURE OF THE DATA

Software are the tools we develop:

- Information systems implemented in Java for the backend and Angular or the GWT framework for the frontend. The data layer relies on PostgreSQL and MySQL relational databases and an Elasticsearch cluster (document-oriented NoSQL).
- Genome analysis tools implemented in Python and C++.

Workflows are:

- ETL tools used to supply data into our IS. Scripts are developed in Bash, Perl or Python and Talend Open Studio projects.
- Analysis tool pipelines, such as the REPET suite for annotating transposable elements, developed with SnakeMake (workflow manager) and Python.

Services are RESTful web services, such as our implementation of the Breeding API - BrAPI, developed in Java.

DATA FORMAT

Software

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

Workflow

- JSON
- SQL
- Bash
- Perl
- Python

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

DATA THEMATIC PERIMETER
- Information management
- Omics

INTELLECTUAL PROPERTY RIGHTS

WHO WILL OWN THE RIGHTS ON THE DATA AND OTHER INFORMATION CREATED?
The institutions that paid for the developments (mainly INRAE) own the codes.

DATA SENSITIVITY

IDENTIFYING THE DATASETS SENSITIVITY LEVEL
- Public
  With the exception of GnpIS (see below), all our code are open and available in version management software. The vast majority of these codes are publicly accessible, but the main branches of their repositories are restricted for certain actions. Others are available on request only.
- Limited distribution
  GnpIS historical database model, GnpIS-coreDB, is private. All codes linked to the structure of this database are therefore restricted to a limited number of users (IS backend, ETL).

WHAT MEASURES ARE TAKEN AND WHAT STANDARDS MUST BE MET TO GUARANTEE THE SECURITY OF SENSITIVE DATA?
The code of our software and ETL tools are managed in version management software, more specifically on forgemia projects, which is the GitLab forge of INRAE's MATHNUM department. We use the application's functionalities to manage access and rights to our repositories.

IF PERSONAL DATA IS INVOLVED, WHAT MEASURES ARE PLANNED TO PROTECT IT DURING THE PROJECT OR IN THE EVENT OF REUSE?
Our codes do not contain any personal data.

DATA SHARING

IS THERE AN OBLIGATION TO SHARE (OR A PROHIBITION OR RESTRICTION)?
There is no obligation to share our code, but we encourage users to get involved (e.g. by giving us feedback, identifying bugs) and developers to improve the code (e.g. by fixing bugs, adding new features) in order to enrich the original GIT repository for everyone to enjoy.
However, sharing GnpIS' historical database model, GnpIS-coreDB, is prohibited.

WHAT ARE THE POTENTIAL REUSES OF THIS DATA?
- Creation of a new instance dedicated to a specific community
- Improvement of existing functionalities
- Addition of new functionalities to meet new needs

DOES DATA READING REQUIRE THE USE OF SPECIFIC SOFTWARE OR TOOLS?
IF SO, WHICH ONE?
ETL tools developed with Talend Open Studio must be managed with the dedicated tool (Talend Open Studio for Data Integration, available free of charge).
Otherwise, no specific software or tools are required to read the code.

HOW DATA WILL BE SHARED?
- Code for GnpIS and its ETLs are versioned on private repositories in the GIT forge forgemia and are available on request only.
- Code for federated portals and their ETLs are versioned on public repositories in the GIT forge forgemia (see Data Discovery, FAIDARE, ETL data portals and RARe-Basket). FAIDARE repository is also mirrored on GitHub (see FAIDARE).

- The Trait Ontology widget and the ETL used to populate FAIDARE with data are publicly available on GitHub (see Trait Ontology widget and ETL FAIDARE).
- Genome analysis tools are versioned GIT forgemia repositories (see repet_pipe, TE_finder et Caulifinder). Some codes are also versioned on public repositories on GitHub (see TE_finder).

WITH WHO?

- Our ETLs and software are shared with anyone (open access).
- GnpIS' historical database model, GnpIS-coreDB, is shared under a proprietary license.

UNDER WHICH LICENCE?

- The code of our federated portals and the ETL used to supply them with data are licensed under the BSD 3-Clause License: redistribution and use in source and binary forms, with or without modification, are permitted under specific conditions.
- GnpIS ETLs are licensed under the GNU LGPL v3 Licence : everyone is authorized to copy and distribute verbatim copies, but modification is forbidden.
- GnpIS' historical database model, GnpIS-coreDB, is licensed under a proprietary license and protected by deposits at the Program Protection Agency.
- REPET tools are licensed under the CeCILL v2.1 Licence: grants users the right to copy, modify and distribute the software governed by this license under an open source distribution model.
- Caulifinder is licensed under the MIT licence: requires preservation of copyright and license notices only.

DATA ORGANISATION AND DOCUMENTATION

WHAT METHODS AND TOOLS ARE USED TO ACQUIRE AND PROCESS DATA, FROM ACQUISITION TO AVAILABILITY, ARCHIVING OR DESTRUCTION?

We use the forgemia GitLab to manage our codes.
With the exception of Talend Open Studio ETLs, which are versioned as they are in the main branch (master) of the repository, any code modification leads to the creation of a dedicated secondary branch (feature). The feature branch is created from the master branch, and any changes to the code are committed to it until the code is ready for production. At this point, the feature branch is merged into the master branch via a merge request, which involves a code review and successful completion of a test suite.

WHAT METADATA WILL BE USED TO ACCOMPANY THE DATASET?
WHAT STANDARDS, VOCABULARIES, TAXONOMIES... WILL BE USED TO DESCRIBE AND REPRESENT DATA AND METADATA?
HOW WILL METADATA BE PRODUCED AND UPDATED?

- The data model used in Data Discovery (the generic code on which the federated portals are based) is derived from the generic data model defined collectively in Spannagl et al. 2016.
- FAIDARE is based on the Breeding API (BrAPI) specifications, itself based on the Minimal Informations About Plant Phenotyping Experiments (MIAPPE) and the Multi-Crop Passport Descriptors (MCPD). It is updated in line with the process set up by the BrAPI community.
- The Trait Ontology widget is based on the BrAPI's "observation variable" web services and is updated according to the process set up by the BrAPI community.

IS ADDITIONAL DOCUMENTATION REQUIRED TO DESCRIBE DATA AND ENSURE LONG TERM REUSABILITY?

README files are available in each software repository to explain how to contribute, install prerequisites, build and run the application, install the Continuous Integration (CI) and set up the configuration and authentication.
Tutorials explaining how to use our analysis tools and documentation explaining how to join our federative portals (see How to join the Data-Discovery federation) are also available alongside the code.
For the ETL tools, a combination of versioned README files in the code repository and internal good practice guides stored on our Wiki or SharePoint (restricted access) are available.

HOW DATA FILES ARE MANAGED AND ORGANISED: VERSION CONTROL, FILE NAMING CONVENTIONS, FILE ORGANISATION

The Trait Ontology widget and the ETL used to supply data to FAIDARE are versioned on GitHub (see ETL FAIDARE et widget Trait Ontology). Other codes are versioned on the forgemia GitLab.
Files explaining how to contribute to the code are also available alongside the code. They describe how to manage the code it GIT, how to manage the application's data, how to set up the development environment and how to run tests (see Contributing to Data-Discovery).

WHAT IS THE DATA QUALITY CONTROL PROCESS?
As explain in the contributing files, each code modification (e.g. bug fixes, new feature, version change) must be performed in a new dedicated branch of the GIT repository. When the branch is ready to be merged into the main branch of the repository, a merge request must be created. At least one core committer from our team reviews the code before validation, and a series of tests is launched and must succeed before the branch is merged and the code released in production.
For our software that are put into production via continuous delivery, i.e. each modification to the master branch results in the code being put into production (as in the case of FAIDARE, WheatIS Data Discovery, RARe and BRC4Env, RARe-Basket), continuous integration is automatically launched by GitLab each time a commit or merge-request is created (see the .gitlab-ci.yml file in Data Discovery).

DATA STORAGE AND SECURITY

HAVE THE STRUCTURE'S INFORMATION SYSTEMS BEEN SUBJECTED TO RISK ANALYSIS OR CERTIFICATION?
Our information systems are not yet certified, but a risk analysis process is underway as part of the implementation of INRAE's IS security policy.

WHAT TYPES OF PHYSICAL MEDIA ARE USED TO STORE DATA?
Our code are versioned on GIT forges: forgemia's GitLab and/or GitHub.

WHAT SECURITY MEASURES ARE IN PLACE FOR DATA TRANSFER?
Data versioned in forgemia can be retrieve using secure HTTPS or SSH protocols. For GitHub, data is transferred via the secure HTTPS protocol.

WHAT IS THE CURRENT AND PROJECTED DATA VOLUME?
- Information Systems: 12,5 GB
- ETLs: 1 GB
- Genome analysis tools: 10 GB

DOES THE ENTITY PHYSICALLY HOSTING THE DATA HAVE AN INFORMATION SECURITY POLICY AND A SECURITY ASSURANCE PLAN?
Forgemia is installed on a VM hosted in INRAE's Toulouse data center and will evolve to become INRAE's institutional forge.
To guarantee high availability and robustness, the service is virtualized on a server hosted in the "L'ARCHE DE DONNÉES Francis Sevila" data center located in INRAE's Occitanie-Toulouse center. Server administration, daily backups and GitLab management are handled by a team of system administrators from various units of INRAE's MATHNUM department, in collaboration with the IT team of Toulouse center.

SECURITY - CONFIDENTIALITY: IS DATA EXCHANGED OR SHARED WITH THIRD PARTIES AND HOW? HOW ARE DATA ACCESS RIGHTS DETERMINED BEFORE PUBLICATION?
Code versioned on public repositories are freely accessible.
Code versioned on private repositories are accessible via LDAP authentication. Access and rights management on these repositories is under our control, thanks to the functionalities offered by GitLab.

SECURITY - INTEGRITY - TRACABILITY: WHAT SAFEGUARDS ARE IN PLACE TO MONITOR DATA PRODUCTION AND ANALYSIS?
We use the functionalities of a GIT forge (forgemia's GitLab and/or GitHub) to manage our codes.

HAVE THE STRUCTURE'S AGENTS BEEN MADE AWARE OF GOOD DIGITAL HYGIENE PRACTICES?
Yes.
All staff have received communications from the INRAE SSI unit. Some front-line staff have also received training in offensive and defensive computer security, and are in contact with the INRAE SSI chain.

ARCHIVING AND DATA RETENTION

WHICH DATA SHOULD BE KEPT FOR THE MEDIUM OR LONG TERM AND WHICH SHOULD BE DESTROYED?
Our codes are to be kept for the long term and will not be destroyed.

ON WHICH PERMANENT ARCHIVING PLATFORM DATA WILL BE STORED FOR LONG TERM RETENTION? WHAT PROCEDURES WILL BE IMPLEMENTED FOR LONG TERM PRESERVATION?

Public GitLab repositories have been synchronized with Software Heritage (an initiative to preserve and share public code source) to enable automatic archiving of our codes at regular intervals (see Data-Discovery, FAIDARE, ETL_data_portals, RARe-Basket, TE_finder and event_caulifinder).

Public GitHub repositories were archived on Software Heritage manually and only once (see Trait Ontology widget, ETL FAIDARE). A more permanent mechanism needs to be put in place (along the same lines as the GitLab system).

HOW LONG WILL THE DATA BE KEPT?

As our codes are versioned on GIT forges (forgemia's GitLab and/or GitHub) they will be accessible for as long as these forges exist. An alternative solution will be used in the unlikely event of these solutions are shut down.

WHAT FUNDING GUARANTEES WILL COVER THE COSTS ASSOCIATED WITH LONG-TERM PRESERVATION?

NA. The code provision service offered by the platform is free of charge. Forgemia's costs are currently supported by INRAE's MATHNUM department and will be directly supported by INRAE when this GIT becomes the institutional forge.

MANAGEMENT MODE « GENOME ANALYSIS »

DATA GENERAL PRESENTATION

DATA OBTENTION MODE

- Data produced by a third party
- Data generated by the structure: Genome annotations (gene nad mobile elements from automatic prediction and/or curation)

ORIGIN

- Analyse

Data for annotating transposable elements are generated using REPET and PanREPET (pangenomic) tools.
Data produced for the annotation of endogenous viruses are generated with the Caulifinder tool.

DATA TYPE

- Dataset
- Service
- Software (described in dedicated management mode)

NATURE OF THE DATA

Input data are assembled genome sequences.
Output data are annotations of mobile elements from automatic predictions and/or curations.

DATA FORMAT

- CSV/TSV
- GFF
- Fasta

DATA THEMATIC PERIMETER

- Omics

INTELLECTUAL PROPERTY RIGHTS

WHO WILL OWN THE RIGHTS ON THE DATA AND OTHER INFORMATION CREATED?

- Datasets generated by URGI members are the property of INRAE. However, URGI members are responsible for the valorization of their datasets.
- Datasets generated by the platform as part of projects must follow the DMP and consortium agreements defined for the project.
- Datasets generated as part of the annotation service offered by the platform are the exclusive property of the service applicants. They will also have full control over their valorization.

In all cases, the platform must be cited in the acknowledgements of any publication associated with the results obtained as follows: *"This work was performed with the facilities of the Plant Bioinformatics Facility (https://doi.org/10.15454/1.5572414581735654E12)"*. URGI members must also use the dual URGI/PlantBioinfoPF affiliation when publishing scientific articles. These requirements are detailed on our website, in the dedicated page « How to cite ».

DATA SENSITIVITY

IDENTIFYING THE DATASETS SENSITIVITY LEVEL

- Limited distribution
  Unless otherwise stated, all data managed during genome analysis are private, as it is subject to specific consortium agreements, originates from private partners and/or is embargoed before publication.
- Confidentiality
  Personal data subject to CNIL and RGPD requirements are confidential.
  On the platform, this data essentially concerns names and emails provided when subscribing to an Offer Of Service (ODS). This personal information is only used by members of the platform to process the request, and is not shared with third parties. It is kept in the best conditions of security and confidentiality, and archived in URGI's project management tool (JIRA) for the life of the scientific data, as defined in this DMP.

In accordance with the European regulation on the protection of personal data (European Regulation 2016/679), the owner of personal data has the right to access, rectify, oppose and delete information concerning him or her. The platform is assisted by INRAE's personal data protection delegate (DPO).

WHAT MEASURES ARE TAKEN AND WHAT STANDARDS MUST BE MET TO GUARANTEE THE SECURITY OF SENSITIVE DATA?
Genome analysis are carried out on Virtual Research Environments (VREs) via individual or clustered Virtual Machines (VMs). A declared SSH key must be supplied to the platform's support team in order to access the VM made available for analysis, and on which the user logs in with a generic centos account. As the owner of this VM, he or she can request access to it by another user, through the platform's support team.
The VM is instantiated on INRAE servers hosted in the Ile-de-France data center.

IF PERSONAL DATA IS INVOLVED, WHAT MEASURES ARE PLANNED TO PROTECT IT DURING THE PROJECT OR IN THE EVENT OF REUSE?
The JIRA instance used by URGI to manage its projects is installed on INRAE servers hosted in the Ile-de-France data center. Access to the confidential data it contains requires authentication, with accounts created on request by our internal services.

DATA SHARING

IS THERE AN OBLIGATION TO SHARE (OR A PROHIBITION OR RESTRICTION)?
There is no obligation to share. The owner of the data, or the person who initiated the request, is responsible for exploiting the data generated. In particular, they must be vigilant in complying with the laws governing research data (open data, regulations on personal data).
However, URGI members are strongly encouraged and supported by the platform to associate DOIs with their datasets and to deposit them in open access in the URGI dataverse on the French open data portal Recherche Data Gouv.
The platform must also be cited in the acknowledgements of any publication associated with the results obtained as follows: *"This work was performed with the facilities of the Plant Bioinformatics Facility (https://doi.org/10.15454/1.5572414581735654E12)"*, as specified in our dedicated page « How to cite ».

WHAT ARE THE POTENTIAL REUSES OF THIS DATA?
- New analysis and/or further analysis related to transposable elements (e.g. impact of TEs on gene expression)
- At the end of each transposable element annotation service, the platform asks the owner if the datasets produced:
    - can be reused for internal URGI analysis
    - can be made available to the community in open access via RepetDB (an information system dedicated to transposable elements) and/or Recherche Data Gouv

DOES DATA READING REQUIRE THE USE OF SPECIFIC SOFTWARE OR TOOLS?
IF SO, WHICH ONE?
No. Data is generated in various open text formats (CSV/TSV, GFF ...).

HOW DATA WILL BE SHARED?
Data available on the Virtual Research Environment is managed by its owner. If the service requester wishes to share his data directly on the VM with someone else, he must ask the platform support to add the new user's SSH key to this VM. Otherwise, data can be uploaded and shared outside the VRE under the responsibility of the service requester.

WITH WHO?
- Identified partner(s)

Users have full control over their data, so they are responsible for sharing and valorizing their datasets. However, there is a validation step on our side before adding a new SSH key to a VM belonging to another user.

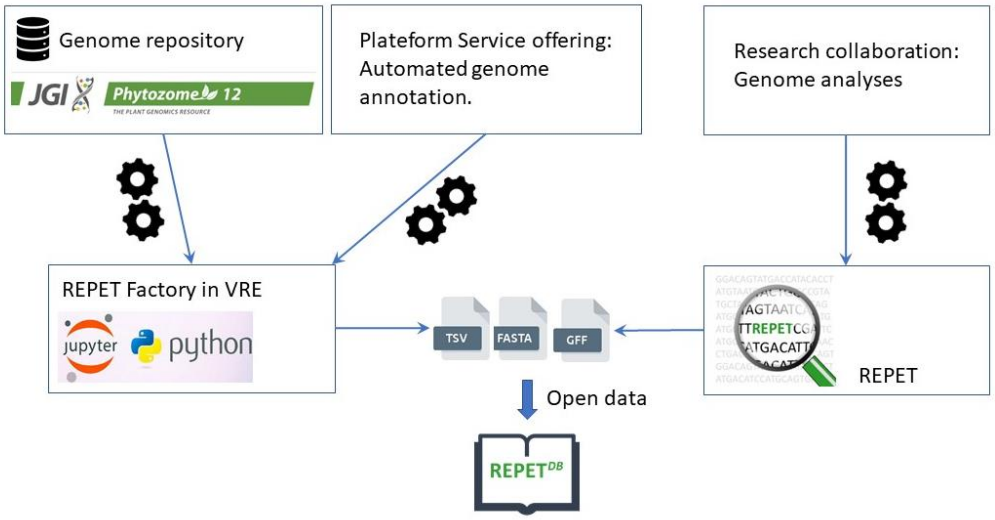
UNDER WHICH LICENCE?
Users have full control over their data, so they are responsible for sharing and valorizing their datasets. Nevertheless, data produced by public research is intended to be published under open licenses (LRPN 2016), and we can give them advice and support in their efforts.

DATA ORGANISATION AND DOCUMENTATION

WHAT METHODS AND TOOLS ARE USED TO ACQUIRE AND PROCESS DATA, FROM ACQUISITION TO AVAIBILITY, ARCHIVING OR DESTRUCTION?

The platform provides the VRE on its OpenStack/Proxmox infrastructure, and handles installation and configuration. The REPET package for detecting and annotating transposable elements in a genome, along with its dependencies, is installed by default on this server.
Users have administrator access to their VRE, so are free to install any additional software they need, provided they comply with our user charter. However, the platform will only provide support for the software it has installed.

Data on transposable elements are generated and managed according to the diagram below (https://urgi.versailles.inrae.fr/files/gnpis/data_integration_repet.jpg).



Consensus libraries of transposable elements are produced by REPET or REPET factory in VREs. Open data are inserted into RepetDB using standard formats: classification file (TSV), REPET Fasta and GFF output. The data is then made accessible via RepetDB, an InterMine-based information system.

WHAT METADATA WILL BE USED TO ACCOMPANY THE DATASET?
WHAT STANDARDS, VOCABULARIES, TAXONOMIES... WILL BE USED TO DESCRIBE AND REPRESENT DATA AND METADATA?
HOW WILL METADATA BE PRODUCED AND UPDATED?

Metadata	**Metadata origin and production mode** *(e.g. manual input, automatic annotation)*	**Standard and associated vocabularies**	**Conditions or frequency update (if applicable)** *(e.g. change in accessibility)*
Assembled genomes	EBI Phytosome	Genome assembly version in fasta format	Fixed standard (fasta) Data updated with each new version
Genome annotations	Phytosome	Genome annotation version in GFF3 format	Fixed standard (gff3) Data updated with each new version
Taxonomy	NCBI	Reference taxonomy	Fixed by NCBI

IS ADDITIONAL DOCUMENTATION REQUIRED TO DESCRIBE DATA AND ENSURE LONG TERM REUSABILITY?

The REPET suite includes tutorials explaining how to use the tool, which are available with the source code (https://urgi.versailles.inrae.fr/Tools/REPET).
Caulifinder documentation is available on a dedicated page of the platform's website: https://urgi.versailles.inrae.fr/Tools/Caulifinder/CAULIFINDER.

HOW DATA FILES ARE MANAGED AND ORGANISED: VERSION CONTROL, FILE NAMING CONVENTIONS, FILE ORGANISATION
VRE data are managed by the owner.

WHAT IS THE DATA QUALITY CONTROL PROCESS?
The REPET suite has its own quality control steps (see https://urgi.versailles.inrae.fr/Tools/REPET).

DATA STORAGE AND SECURITY

HAVE THE STRUCTURE'S INFORMATION SYSTEMS BEEN SUBJECTED TO RISK ANALYSIS OR CERTIFICATION?
Our information systems are not yet certified, but a risk analysis process is underway as part of the implementation of INRAE's IS security policy.

WHAT TYPES OF PHYSICAL MEDIA ARE USED TO STORE DATA?
Data managed on the VRE (copied and/or generated data) are stored on CEPH volumes mounted on the VM. Data generated at all REPET stages are managed in a dedicated MySQL database hosted on our OpenStack/Proxmox infrastructure. Everything is hosted in INRAE's Ile-de-France data center.

WHAT SECURITY MEASURES ARE IN PLACE FOR DATA TRANSFER?
The relevant data is copied to the VRE using cryptographic tools, such as secure shell sessions (SSH).

WHAT IS THE CURRENT AND PROJECTED DATA VOLUME?
The total volume available is around 70 TB in July 2021.
It can be easily increased if required (the infrastructure is easily adaptable and the necessary resources are available).

DOES THE ENTITY PHYSICALLY HOSTING THE DATA HAVE AN INFORMATION SECURITY POLICY AND A SECURITY ASSURANCE PLAN?
Our security policy follows the INRAE research infrastructure charter.
Data managed on the VRE are stored on URGI servers, hosted in the Ile-de-France data center with strict physical access control. CEPH volumes and the OpenStack/Proxmox VM are not backed up at this stage, but the data exists in three copies on CEPH and the rebuilding of bare VMs for the VRE is automated.

SECURITY - CONFIDENTIALITY: IS DATA EXCHANGED OR SHARED WITH THIRD PARTIES AND HOW? HOW ARE DATA ACCESS RIGHTS DETERMINED BEFORE PUBLICATION?
A declared SSH key must be provided to access the VRE made available by the platform. Only the service subscriber can request to open access to his or her VRE and its data to another user.
Data can be accessed using cryptographic tools, such as secure shell sessions (SSH) to copy files. Access to these files is controlled using well-known UNIX permissions on Linux servers, CEPH data volumes are only accessible on the user's VM and cannot be mounted elsewhere.

SECURITY - INTEGRITY - TRACABILITY: WHAT SAFEGUARDS ARE IN PLACE TO MONITOR DATA PRODUCTION AND ANALYSIS?
VMs on URGI's OpenStack/Promox are supported by CEPH technology, and benefit from its own integrity and security features such as scrubbing (triple replication of data on different physical storages, regular comparison of replicated data) or redundancy cycle checks (CRC recalculating the footprint of data as it is read and comparing it with that stored when it was written).

HAVE THE STRUCTURE'S AGENTS BEEN MADE AWARE OF GOOD DIGITAL HYGIENE PRACTICES?
Yes.
All staff have received communications from the INRAE SSI unit. Some front-line staff have also received training in offensive and defensive computer security, and are in contact with the INRAE SSI chain.

ARCHIVING AND DATA RETENTION

WHICH DATA SHOULD BE KEPT FOR THE MEDIUM OR LONG TERM AND WHICH SHOULD BE DESTROYED?
At the end of the resource allocation period, the VRE is closed and the data erased within 15 days. As there is no archiving, users must retrieve the data they wish to keep before this deadline.
For URGI members, the VRE and its data are kept for as long as necessary on a mounted and backed-up storage space.

ON WHICH PERMANENT ARCHIVING PLATFORM DATA WILL BE STORED FOR LONG TERM RETENTION? WHAT PROCEDURES WILL BE IMPLEMENTED FOR LONG TERM PRESERVATION?

There is no data archiving action. However, we are proposing publication of the datasets in RepetDB (an information system dedicated to transposable elements) and/or the URGI dataverse of Recherche Data Gouv portal, the latter guaranteeing access to the data for 10 years thanks to the attribution of a DOI.

HOW LONG WILL THE DATA BE KEPT?

At the end of the resource allocation period, users have 15 days to retrieve the data they wish to keep.

WHAT FUNDING GUARANTEES WILL COVER THE COSTS ASSOCIATED WITH LONG-TERM PRESERVATION?

INRAE provides core funding (human resources and financial support for the platform) via the BAP and ECODIV research departments and the CNOC (“Commission Nationale des Outils Collectifs”).
The platform also has access to specific funding (paying ODS).

GLOSSARY

ANR	« Agence Nationale de la Recherche »
API (RESTful)	Application Programing Interface (REpresentational State Transfer)
BAM	Binary Alignment Map (file format)
BAP	« Biologie et Amélioration des Plantes », INRAE's department on Plant Biology and Breeding (https://www.inrae.fr/departements/bap)
BED	Browser Extensible Data (file format)
BrAPI	Breeding API (https://brapi.org/)
CNIL	« Commission Nationale de l'Informatique et des Libertés » (https://www.cnil.fr/en)
CSV	Coma Separated Values (file format)
DBMS	DataBase Management System
DMP	Data Management Plan
DOI	Digital Object Identifier
DPO	Data Protection Officer
ECODIV	« Écologie et Biodiversité », INRAE's department on Biology and Biodiversity (https://www.inrae.fr/departements/ecodiv)
ETL	Extract Transform Load
FP7	7th Framework Programme for Research
GFF	General feature format (file format)
GDPR	General Data Protection Regulation (https://gdpr-info.eu/)
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 », the French Institute of Bioinfomatics (https://www.france-bioinformatique.fr/en/home/)
INRAE	« Institut National de Recherche pour l'Agriculture, l'Alimentation et l'Environnement », the French National Research Institute for Agriculture, Food and Environment (https://www.inrae.fr/)
IS	Information System
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 (file format)
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)
PIA	« Plan d'Investissement d'Avenir »
QTL	Quantitative Trait Loci
RDA	Research Data Alliance
RDBMS	Relational DataBase Management System
RDF	Resource Description Framework (file format)
SAM	Sequence Alignment Map
SCSS	Syntactically Awesome Style Sheet (file format)
SQL	Structured Query Language (language)
SSH	Secure Socket Shell
TE/ET	Transposable Elements / Eléments Transposables
TSV	Tabulated Separated Values (file format)
URGI	« Unité de Recherche en Génomique-Info », INRAE's Research Unit Genomics-Info (https://www6.versailles-grignon.inrae.fr/urgi/)
VCF	Variant Call Format (file format)
VM	Virtual Machine
VRE	Virtual Research Environment
YAML	Yet Another Markup Language (file format)