



HAL
open science

Les services de FAIRification des données à l'URGI : cas d'utilisation par des projets/consortium nationaux et internationaux

Michaël Alaux, Anne-Françoise Adam-Blondon, Françoise Alfama, Etienne Bardet, Raphaël Flores, Nicolas Francillonne, Vanita Haurheeram, Najwa Lakmouri, Erwan Le Floch, Maud Marty, et al.

► To cite this version:

Michaël Alaux, Anne-Françoise Adam-Blondon, Françoise Alfama, Etienne Bardet, Raphaël Flores, et al.. Les services de FAIRification des données à l'URGI : cas d'utilisation par des projets/consortium nationaux et internationaux. Réunion annuelle de l'IFB, Dec 2025, Paris, France. <hal-05530229>

HAL Id: hal-05530229

<https://hal.science/hal-05530229v1>

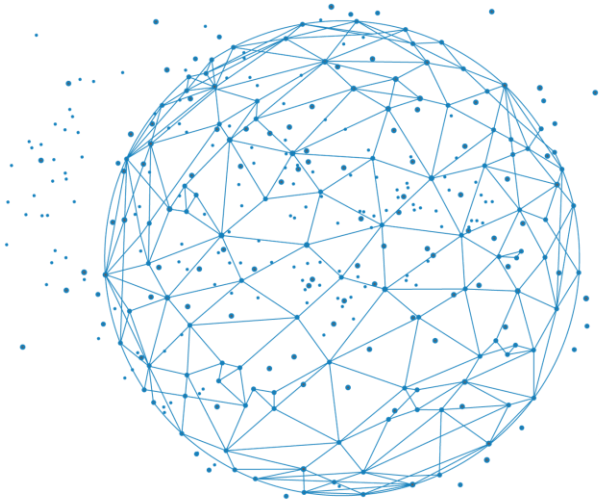
Submitted on 27 Feb 2026

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons CC BY 4.0 - Attribution - International License



Les services de FAIRification des données à l'URGI : cas d'utilisation par des projets/consortium nationaux et internationaux

Michaël Alaux & Anne-Françoise Adam-Blondon





Unité Ressources Génomique-Info

- Unité de service INRAE (centre Versailles-Saclay) en bioinformatique pour les plantes d'intérêt agronomique et forestières
- Membre de la Graduate School Biosphera et du réseau Saclay Plant Science
- Labellisations :
 - INRAE: labellisée ISC et membre de l'IR INRAE BioinfOmics
 - National: labellisée IBiSA et membre de l'IFB (IR nationale)
 - Certifiée ISO 9001
- Forte implication à l'Europe et l'international dans le développement de standards et recommandations pour la gestion des données "plantes" selon les principes FAIR
- Partenaire de nombreux projets nationaux et internationaux pour la gestion de données





Développement d'une science ouverte

Partenariats UMR/UE, DipSO

Plan données pour la science



Contribution aux Infrastructures de Recherche

- ISC
 - IR BioinfOmics
 - Soutien/membre IR : IFB, RARe, Phenome-Emphasis, In-Sylva, France génomique



Services

- Entrepôt de données FAIR
 - Caractérisation de ressources génétiques
 - Données de phénotypage
 - Annotation de génomes
- Fédérations de données
- Annotation des éléments transposables (ET)
 - Ressources de calcul et prestation d'analyse
 - Outils bioinformatiques

Les services développés ou mobilisés autour de la donnée



- Development and maintenance of metadata standards and ontologies for plant phenotyping data (including forests)
- Development of guidelines for data management according to FAIR principles



- Data discovery portals: international (FAIDARE, WheatIS) and national (RARE)
- Contribution to the development of an international web services standard: BrAPI



- GnpIS Information System: INRAE genetic resource data and phenotyping data
- Support for FAIR datasets submission to Recherche Data Gov



Supporting infrastructures and major projects with their data management plan



- Provision of reference data (genetic resources, identifiers/DOI, consensus of transposable elements, etc.)
- Provision of templates for metadata collection

- Provision of a space for sharing and curating metadata (FAIRDOME-SEEK, Nextcloud)
- Metadata curation and enrichment

- Reproducible annotation of repeated elements in genomes (REPET Factory)
- Provision of containerized pipelines: REPET
- Provision of Virtual Research Environments





Browse all topics by



Data life cycle

Start here to get an overview of research data management based on stages in the data life cycle.



Your role

Identify your role in research data management, find data management resources relevant for you, and information to help you progress in your career path.



Your domain

Learn about data management tasks that affect your domain or research community, and the solutions adopted to address them.



Your tasks

Find guidelines and solutions for tackling common data management tasks.



Tool assembly

Find concrete combinations of tools and resources assembled into an ecosystem for research data management.



National resources

Find pointers to country specific information resources and national research data management practices.



All tools and resources

Browse the RDMkit's catalogue of tools and resources for research data management.



All training resources

Browse all training resources mentioned in RDMkit pages.

RDMkit

Linked to other registries:

- ✓ FAIRCookbook
- ✓ FAIRsharing
- ✓ TeSS
- ✓ Bio.tools
- ✓ WorkflowHub
- ✓ Data Stewardship Wizard

<https://rdmkit.elixir-europe.org>



We welcome contributors!

This project would not be possible without the many [amazing community contributors](#). RDMkit is an open community project, and you are welcome to join us!

Start contributing



What's new?

New page: Data deletion

06 November 2025 - 1715

A "your task" page about how to properly delete data was added. [Discover the page here.](#)

New page: Biodiversity

28 August 2025 - 1711

A new "your domain" page about Biodiversity was added. [Discover the page here.](#)

New page: Virology

15 May 2025 - 1620

A new "your domain" page about Virology was added. [Discover the page here.](#)

For more news please visit our [news page](#).

Ex: Contribution récente française IFB – Data Terra/PNDB au développement d'une page de guidelines dédiée au domaine « Biodiversité »

Un outil extrêmement populaire développé par ELIXIR-FR: [FAIR-Checker](#)



RDMkit in numbers

231

Contributors

The force behind RDMkit



599

Tools & resources

Explained in the context of real world problems



136

Pages

Helping you with data management





Your domain

Plant sciences

RDMkit

Introduction

Data management challenges in plant sciences

The plant science domain includes studying the adaptation of plants to their environment, with applications ranging from improving crop yield or resistance to environmental conditions to managing forest ecosystems. Data integration and reuse are facilitators for understanding the play between genotype and environment to produce a phenotype, which requires integrating phenotyping experiments and genomic assays made on the same plant material with geo-climatic data. Moreover, cross-species comparisons are often necessary to understand the mechanisms behind phenotypic traits, especially at the genotypic level, due to the gap in genomic knowledge between well-studied plant species (namely Arabidopsis) and newly sequenced ones.

The challenges to data integration stem from the multiple levels of heterogeneity in this domain. It encompasses a variety of species, ranging from model organisms to crop species to wild plants such as forest trees. These often need to be detailed at infra-specific levels (e.g. subspecies, variety), but naming at these levels sometimes lacks consensus. Studies can take place in a diversity of settings, including indoor (e.g. growth chamber, greenhouse) and outdoor settings (e.g. cultivated field, forest) which differ fundamentally on the requirements and manner of characterizing the environment. Phenotypic data can be collected manually or automatically (by sensors and drones), and be very diverse in nature, spanning physical measurements, the results of biochemical assays, and images. Some omics data can be considered as well as molecular phenotypes (e.g. transcriptome, metabolomes). Thus, the extension and depth of metadata required to describe a plant experiment in a FAIR-compliant way is very demanding for researchers.

Another particularity of this domain is the absence of central deposition databases for certain important data types, in particular data deriving from plant phenotyping experiments. Whereas datasets from plant omics experiments are typically deposited in global deposition databases for that type of experiment, those from phenotyping experiments remain in institutional or, at best, national repositories. This makes it difficult to find, access and interconnect plant phenotyping data.

Data management planning

Description

The general principles for data management planning are described in the [Planning](#) page of the Data life cycle section, while generic but more practical aspects of writing a DMP can be found on the [Data Management Plan](#) page.

Considerations

- Important general considerations about data management planning can be found on the [Planning](#) page.
- Phenotyping data must be described following the [MIAPPE](#) data standard.
- Make sure to identify and describe the biological material and the observation variables in your research.

Exemple du domaine « Plantes »

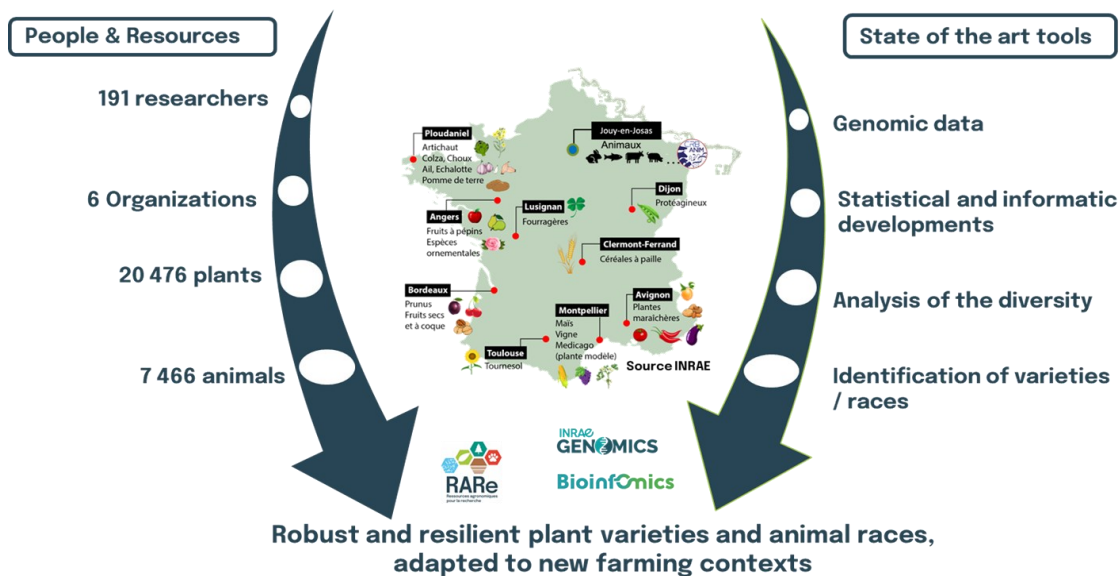
- Ressources support à la gestion FAIR des données sur les plantes développées depuis 2015 dans un cadre international : Wheat Initiative (collaboration RDA), l'ELIXIR Plant Sciences Community, EMPHASIS, BrAPI consortium, MIAPPE consortium, CGIAR
 - Compilation dans la page “Plant Sciences” du RDMkit
- « Crash test » et amélioration dans des projets Internationaux, Européens et nationaux

Solutions



Principaux objectifs de BReIF:

- **Interopérabilité des services numériques** des infrastructures de recherche française : RARe, IFB, France génomique
- Amélioration de la **gestion FAIR** des données

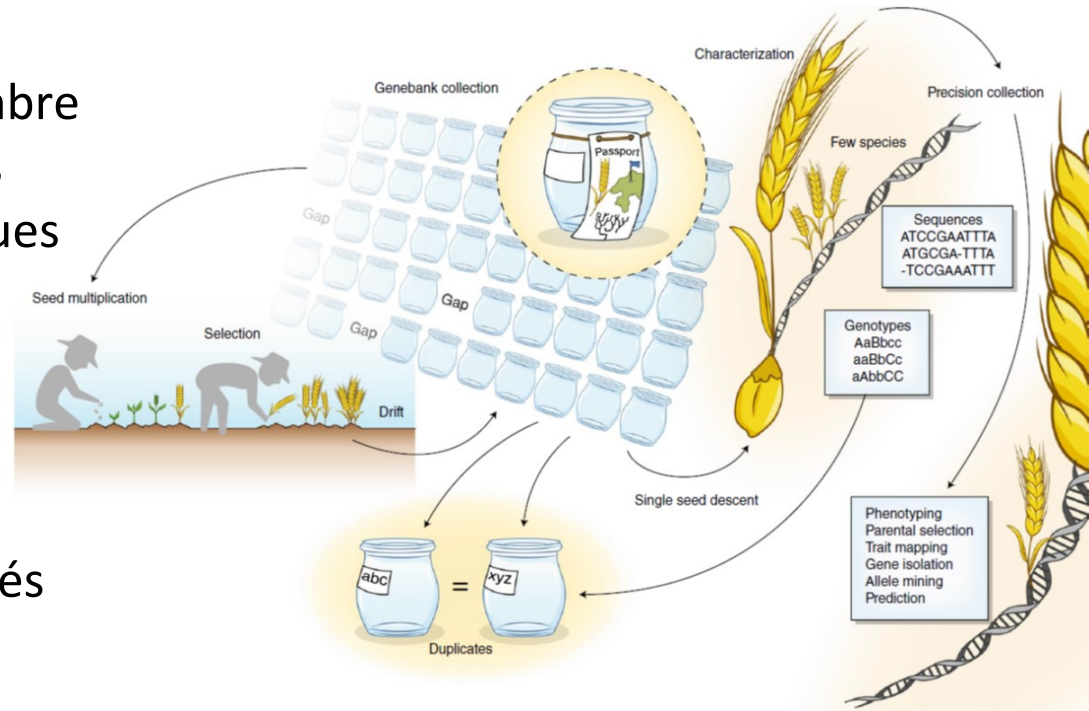




caractérisation d'un grand nombre de collection de blé et orge des centres de ressources biologiques européens



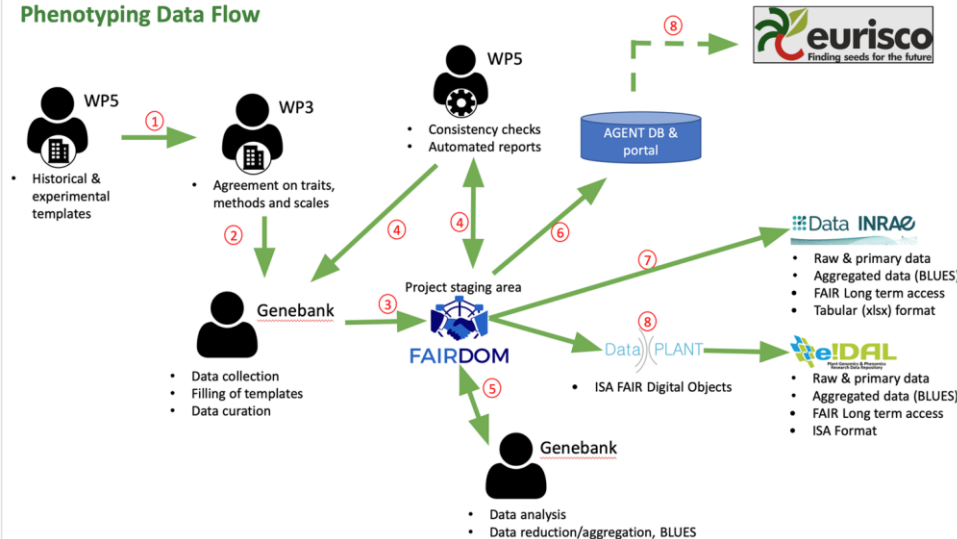
sur espèces sauvages apparentés aux espèces cultivés



- Mise à disposition de guidelines pour gérer les données du projet
 - Complémentaire du plan de gestion de données
 - Flux de données par type de données
 - Formats, standards de données et outils à utiliser
 - Référence aux sources externes (RDMkit, FAIRCookbook)



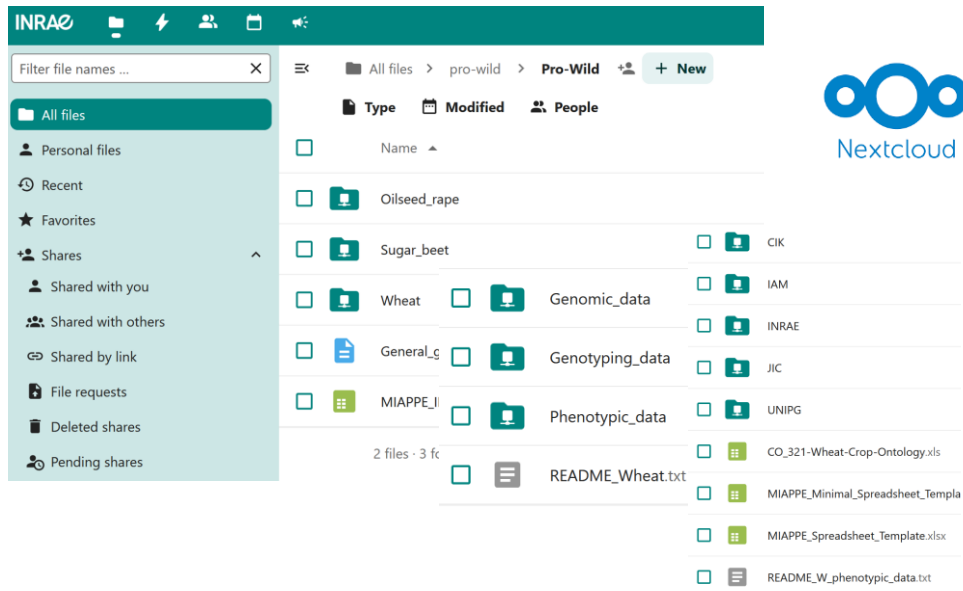
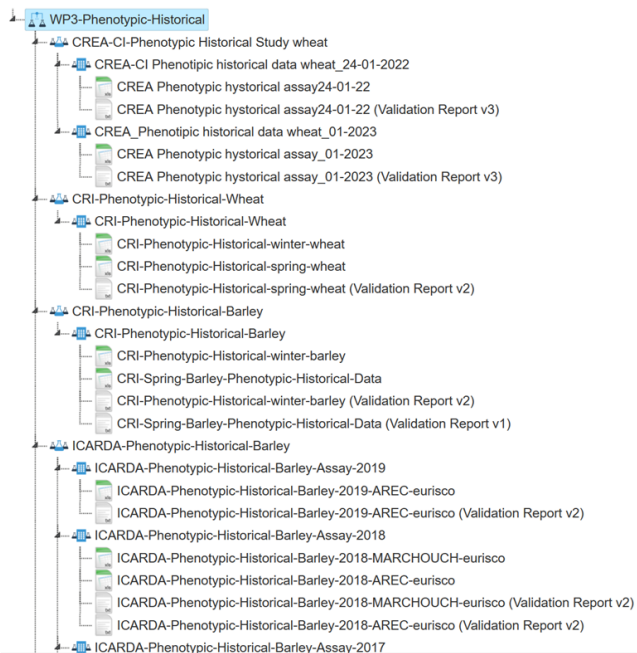
Phenotyping Data Flow



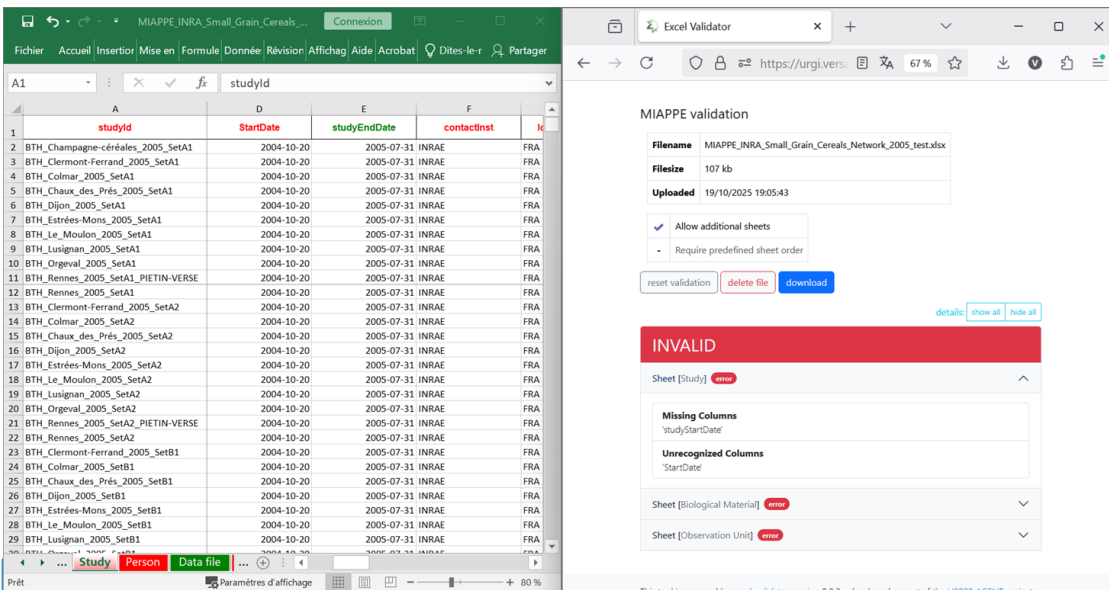
<https://doi.org/10.5281/zenodo.12625359>



- Espace de partage, curation et validation de données
 - arborescence
 - contient templates, fichiers d'exemple, guidelines, README



- Outil d'aide à la curation de métadonnées de phénotypage
 - version en ligne <https://urgi.versailles.inrae.fr/excel-validator/miappe>
 - lancement automatique dès qu'un nouveau fichier est déposé et écriture de rapport



The screenshot shows the Excel Validator interface with an invalid report. The report details are as follows:

Filename	MIAPPE_INRA_Small_Grain_Cereals_Network_2005_test.xlsx
Filesize	107 kb
Uploaded	19/10/2025 19:05:43
<input checked="" type="checkbox"/> Allow additional sheets	
<input type="checkbox"/> Require predefined sheet order	

Buttons: reset validation, delete file, download

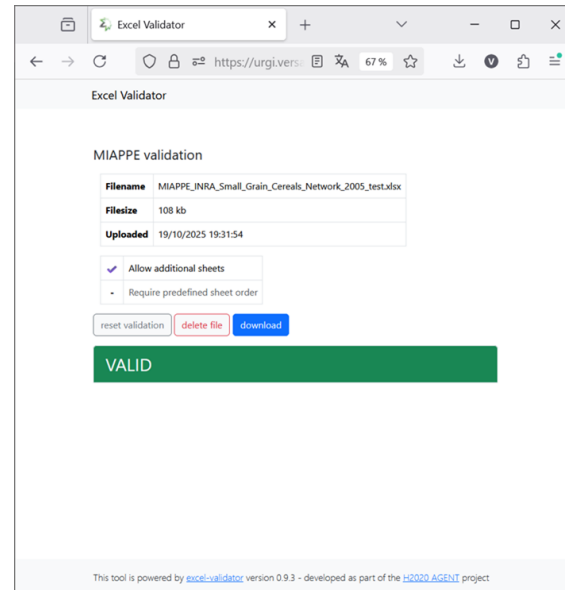
INVALID

Sheet [Study] **error**

- Missing Columns: *studyStartDate
- Unrecognized Columns: *startDate

Sheet [Biological Material] **error**

Sheet [Observation Unit] **error**



The screenshot shows the Excel Validator interface with a valid report. The report details are as follows:

Filename	MIAPPE_INRA_Small_Grain_Cereals_Network_2005_test.xlsx
Filesize	108 kb
Uploaded	19/10/2025 19:31:54
<input checked="" type="checkbox"/> Allow additional sheets	
<input type="checkbox"/> Require predefined sheet order	

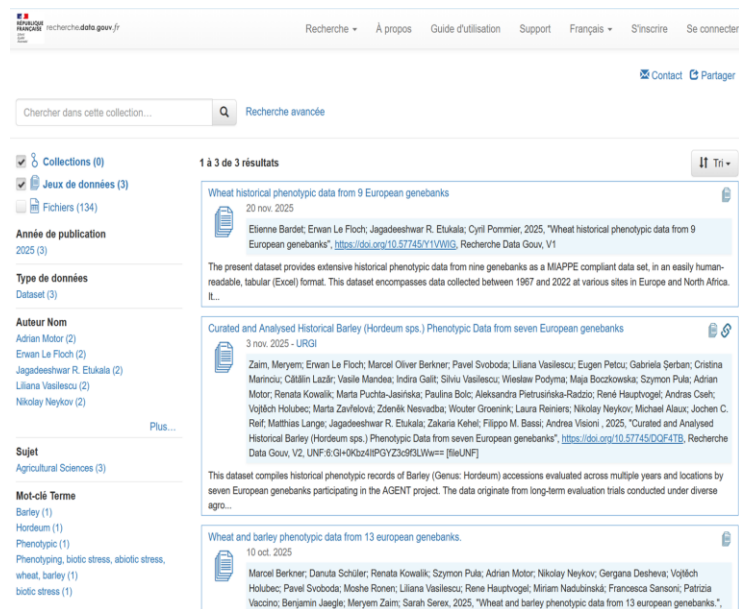
Buttons: reset validation, delete file, download

VALID

Footer: This tool is powered by excel-validator version 0.9.3 - developed as part of the H2020 AGENT project

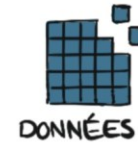
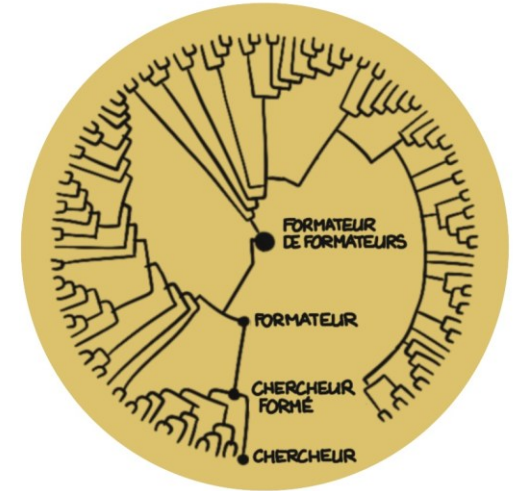
- Disciplinaires: EBI/BioSamples et ENA pour les séquences et échantillons associés
- Recherche Data Gouv: jeux de données de phénotypage curés, validés et analysés
- DataPLANT: FAIR Digital Object (RO-Crate, ARC) → machine actionable

Scripts en accès ouvert pour faciliter la soumission et la récupération d'identifiants



The screenshot displays the Recherche Data Gouv website interface. At the top, there are navigation links for 'Recherche', 'À propos', 'Guide d'utilisation', 'Support', 'Français', 'S'inscrire', and 'Se connecter'. A search bar is present with the text 'Chercher dans cette collection...' and a search icon. Below the search bar, there are filters for 'Collections (0)', 'Jeux de données (3)', and 'Fichiers (134)'. The main content area shows search results for 'Wheat historical phenotypic data from 9 European genebanks' and 'Curated and Analysed Historical Barley (Hordeum sps.) Phenotypic Data from seven European genebanks'. Each result includes a date, a list of authors, and a brief description of the dataset. The website is in French.

- Visibles dans le [catalogue de formation de l'IFB](#) qui alimente automatiquement celui d'ELIXIR ([TeSS](#))
- Du **matériel de formation réutilisable** accessible via une instance [Moodle](#) de l'IFB



Fiabilité



Accessibilité



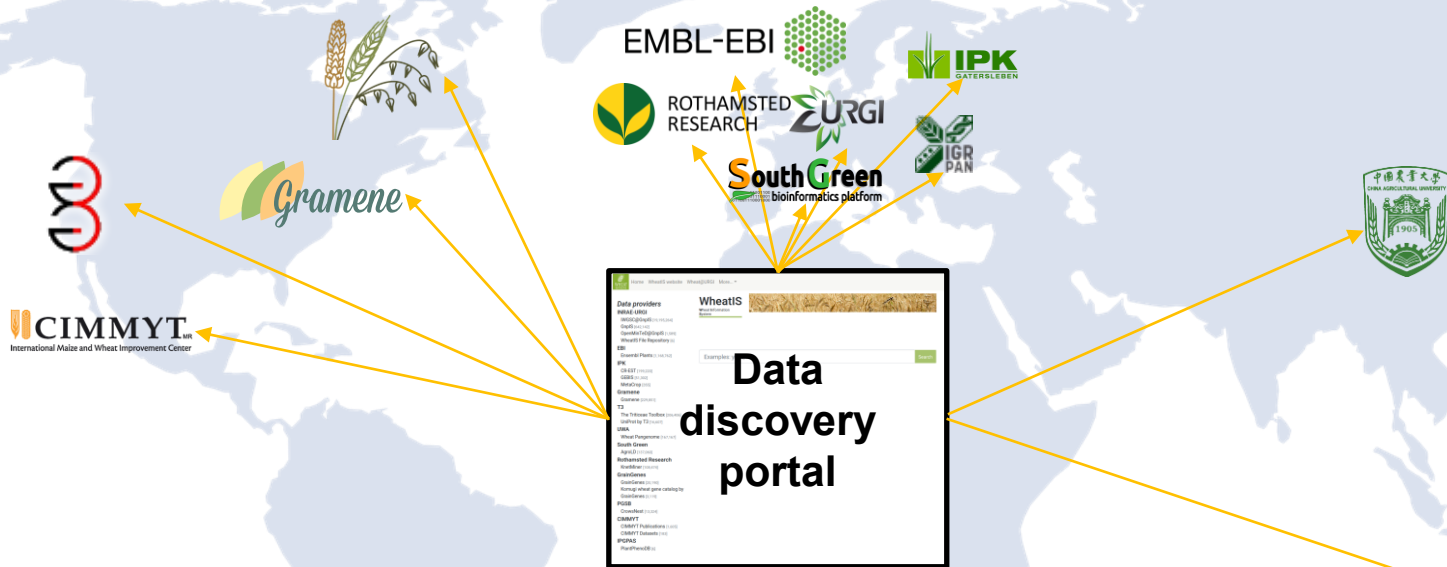
Interopérabilité



Réutilisable

crédit Jordane Ancelin

Portail de recherche des données scientifiques



Fédérations de données partageant les mêmes standards



<https://urgi.versailles.inrae.fr/wheatis/>



<https://urgi.versailles.inrae.fr/aidare/>

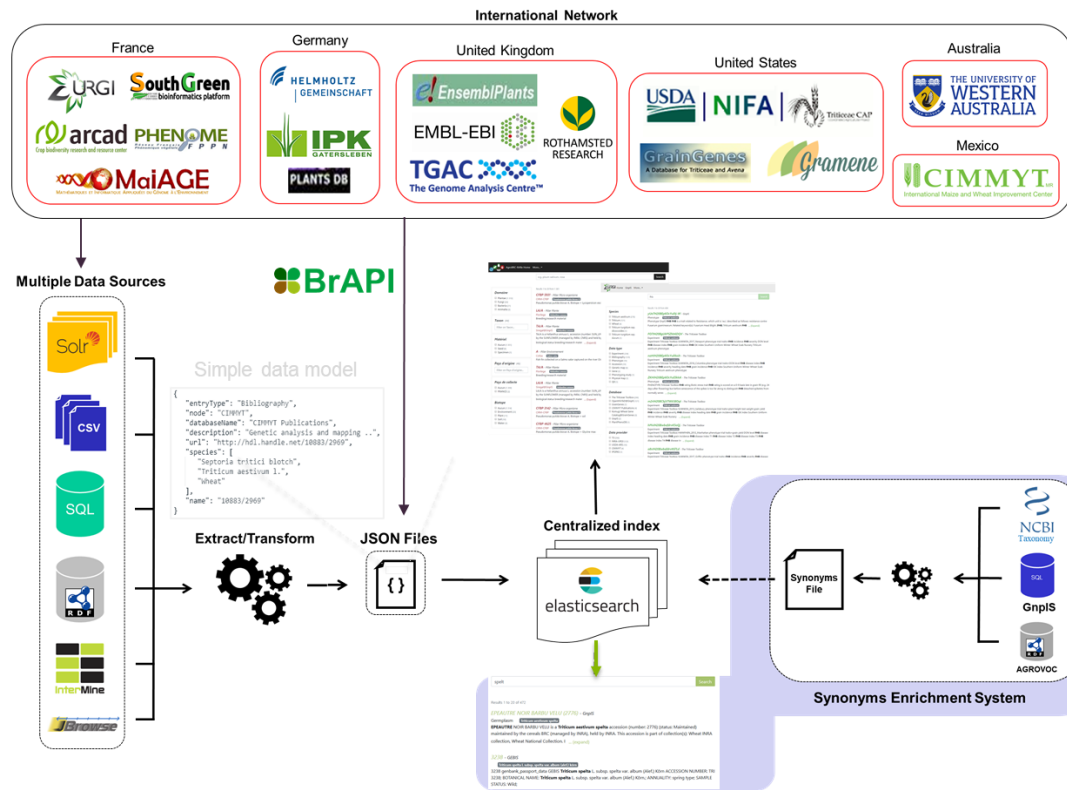




28 bases de données indexées
pointant vers
30 millions données



40 bases de données indexées
pointant vers
97 millions de données



Exemple de recherche

cfid2

Genome Track View Help

Espèces 0 50,000,000 100,000,000

Select tracks

Triticum a

Aegilops f

Triticum a

Cfd SSR

Type de données

HighConfidenceGenesv1.1 (GOMAP annotation)

Gene annotation

Physical map

Marker [2]

Genome map

CLARITE (TE model)

TaBW280K

Base de données

IWGSC@GnplS [8]

GnplS [2]

Fournisseurs de données

INRAE-URGI [10]

[JBrowse_chr4A](#)

Gene annotation
SEQUENCE FEATURE
= ePCR, Seq_id =
[...\(voir la suite\)](#)

[JBrowse_chr5D](#)

Gene annotation
SEQUENCE FEATURE
0, Source = ePCR

gene TraesCS2A02G151900

Right click on this feature to get links to external resources

Primary Data

Name TraesCS2A02G151900 [Search in WheatIS](#)

Type gene

Score 61

Position chr2A:100113981..100117432 (+ strand)

Length 3,452 bp

Attributes

Go (12)

[GO:0001510](#) [GO:0001522](#) [GO:0006164](#) [GO:0006364](#) [GO:0065007](#) [GO:0003723](#)

[GO:0009982](#) [GO:0005654](#) [GO:0005829](#) [GO:0009506](#) [GO:0019013](#) [GO:0031429](#)

Id TraesCS2A02G151900

Previous_id TraesCS2A01G151900

Primconf HC

Seq_id chr2A

Source IWGSC_v1.1_201706

Region sequence

FASTA

```
>chr2A chr2A:100113981..100117432 (+ strand) class=gene
length=3452
GCCACGGCGCCGCTTCCGTCGCCCGCAAAAACCCCTAATCCCCCTCCCCACCCACCCAC
CCCACCCGCGCCATGTCGTCCACCGCCCGCGCGCTCGCGTCCCCGCTCCGAGCACACCA
AATCTAAGAAAAAGAAGACAAATCCAAGGACGCTCCGCGACCCCGCCCGGATACACCA
TCGCTGGCGGAGCCGAGGCCAAGACTGATGGTATATGATCAAGCCCAAGGCCCTGGTTCC
GTCCCTCGACACCTCCACCTGGCCGCTCCTCCTCAAGAACTACGACCGCCTCAAGTCCGCA
CCGGCCACTACACCCCGCTCCCTCCGGCCACTCGCCGCTCAAGCGCCCTCCGCGAGTAC
CTCCGCTACGGCGTCATCAACCTCGACAAGCCGTCACACCCCTCCTCCACGAGGTGGTGGC
GTGGATCAAGCGCTCCTCCGCTCGAGAAGACCGGCCACAGCGGCACGCTCGACCCCAAGG
TCACCGGCAACCTCATCGTCTGCTCGACCGGCCACAGCCTCGTCAAGTCCGACGAGGGC
```

Rechercher

0,000 350,000,000

[-] [-] [+] [+]



3940, Strand = 0, Source stivum similarity chr4

l = 517967269, Strand = iticum aestivum simili



Gene Card

Input gene ID

Concept: pseudouridine synthase
Type: Protein

Concepts: 17 (248); Relations: 19 (324)

Interactive Legend:

0/20	4/14	2/11	0/1	0/9	4/6	0/1	0/94	0/2	0/2	0/70	0/3	0/1	0/5
Publication	Gene	Protein	Pathway	Domain	SNP	SNPEffect	PO	CoExpCluster	CoExpStudy	BioProc	MolFunc	EC	CellComp



Espèces

- Triticum [1]
- Triticum aestivum [1]

Type de données

- Bibliography [1]

Annotation ontologique

- Typhula ishikariensis* (WTO:0000445) [1]
- snow mold (WTO:0000515)
- Étendre la recherche ?

Base de données

- OpenMinTeD@GnpIS [1]

Fournisseurs de données

- INRAE-URGI [1]

Notes

Mapping QTL conferring speckled snow mold resistance in winter wheat (*Triticum aestivum* L.)

Zenta Nishio, Norio Iriki, Miwako Ito, Tadashi Tabiki, Timothy Murray

[+ Author information](#)

Keywords: [winter wheat](#), [snow mold resistance](#), [quantitative trait loci](#), [doubled haploid lines](#)

[JOURNAL](#) [FREE ACCESS](#) [FULL-TEXT HTML](#)

2020 Volume 70 Issue 2 Pages 246-252

[DOI](#) <https://doi.org/10.1270/jsbbs.19111>

[i](#) Browse "Advance online publication" version

[+ Details](#)

Article overview

- [> Abstract](#)
- [> References \(43\)](#)
- [> Figures \(3\)](#)
- [> Content from these authors](#)
- [> Cited by \(4\)](#)

Share

Abstract

Speckled snow mold caused by *Typhula ishikariensis* is one of the most devastating diseases of winter wheat in Hokkaido, Japan and parts of the Pacific Northwest region of USA. Münstertaler is a winter wheat landrace from Switzerland that has very high resistance to snow mold and superior freezing tolerance. Quantitative trait loci (QTL) for resistance to speckled snow mold were identified in a doubled haploid population derived from a cross between Münstertaler and susceptible variety Ibis, both under field conditions and controlled environment tests. Composite interval mapping analysis revealed a

[Full-text HTML](#)

[Download PDF \(510K\)](#)

[Download citation](#)

RIS
(compatible with EndNote, Reference Manager, ProCite, RefWorks)

BIB TEX
(compatible with BibDesk, LaTeX)

Text

[How to download citation](#)

[Contact us](#)

Favorites & Alerts

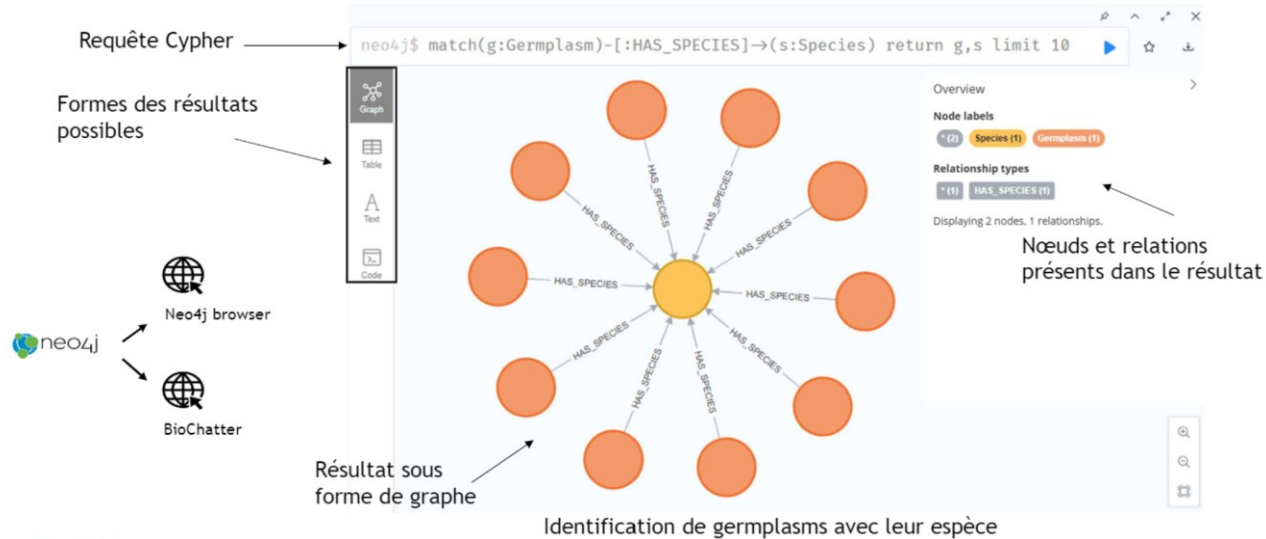
- [☆ Add to favorites](#)
- [☆ Additional info alert](#)
- [☆ Citation alert](#)
- [☆ Authentication alert](#)



[Related articles >](#)

g QTL conferring speckled
osome 5D was about 5 cM

- Intégration de données dans des graphes de connaissances
- Utilisation de l'IA pour améliorer intégration, curation, interface homme-machine



Conclusions de notre retour d'expérience

- Les infrastructures de recherche, IFB et ELIXIR, ont permis de capitaliser sur le long terme les développements de ressources support
 - maintenance des ontologies, standards, recommandations
 - maintenance des entrepôts de données
 - personnels experts dans la standardisation et curation de données (*data stewards*)
- Les ressources support doivent être co-construites avec les communautés de recherche (ex: Wheat Initiative)
 - cadrage des besoins et expertise sur les données
 - personnes relai pour diffuser les bonnes pratiques

Identification des partenaires et des ressources:

- Ecosystème Recherche Data Gouv
- forum des ESFRI

- Impliquer l'infrastructure de recherche pertinente comme partenaire de projet
- Faire travailler les infrastructures ensemble



- Les activités de formation et de montée en capacité sont critiques
 - former les chercheurs
 - former des formateurs
 - développement de matériel de formation réutilisable
- Dans un contexte de grande dispersion des données, le développement d'une **féderation de données** facilite la mise en place de réseaux de collaborations entre bases de données

Identification de partenaires et de ressources:

- Ecosystème Recherche Data Gouv
- forum des ESFRI

Les systèmes de labellisation des bases de données partenaires devraient permettre d'identifier les forces et faiblesses pour consolider ces fédérations



Raphaël Flores : coordination opérationnelle des services

Cyril Pommier : co-pilote services données et fédération, responsable communauté plante d'ELIXIR, responsable données Phenome-Emphasis

Célia Michotey : co-pilote services données et fédération

Vanita Haurheeram : data steward BReIF et Pro-Wild

Erwan Le-Floch et Etienne Bardet : data steward AGENT

URGI team





Michaël Alaux : michael.alaux@inrae.fr
Anne-Françoise Adam-Blondon :
anne-francoise.adam-blondon@inrae.fr
<https://urgi.versailles.inrae.fr/>



INSTITUT FRANÇAIS DE BIOINFORMATIQUE

