



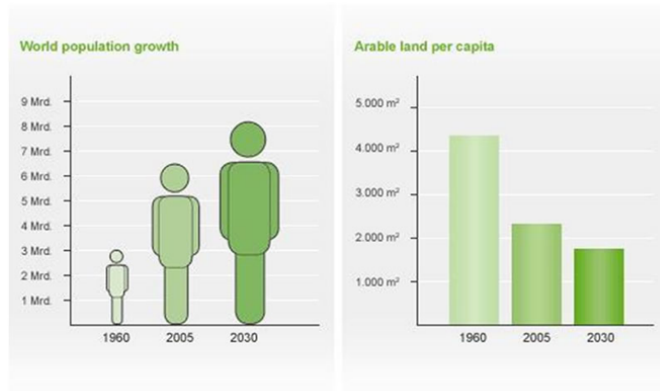
Heterogenous plant data from data integration practices to analysis perspectives.

Intégration de données hétérogènes de plantes : des pratiques d'intégration aux perspectives d'analyse



Motivation: Increased sustainability of plant production is needed

- Less arable land
- Increasing world population
- Climate change (extreme weather conditions, pests and diseases)



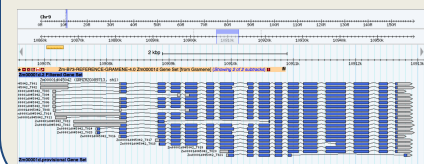
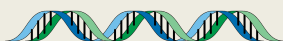
Source: FAO



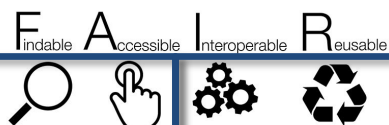
12 of the 17 UN Sustainable Development Goals (SDGs) address agriculture directly and indirectly

Genome / Environment / Phenome

Genetics & Genomics



Homogenous (Mostly)
Central Access (Mostly)



Phenomics



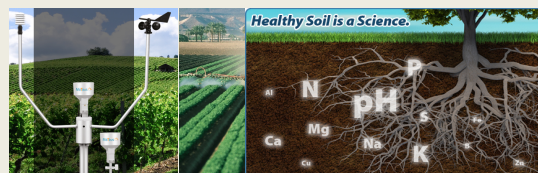
Heterogenous
Distributed

Plant Breeding
Genetic variations by Traits

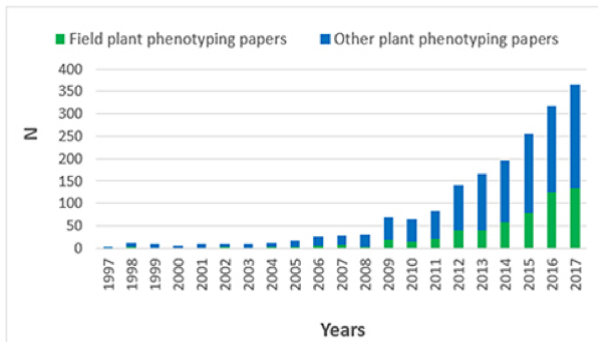
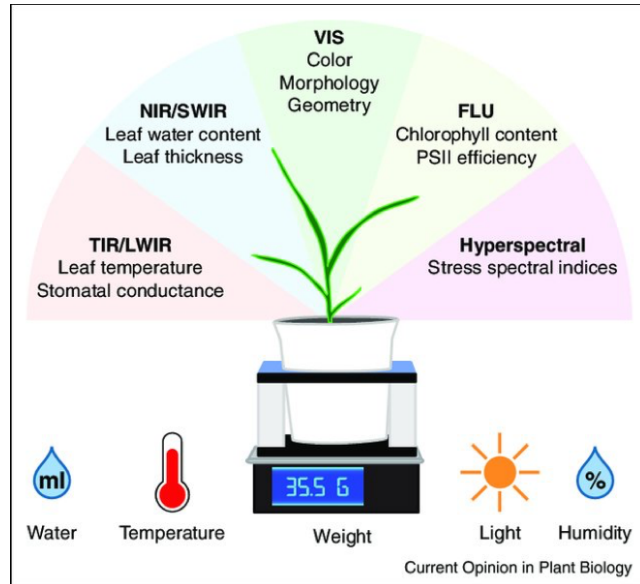


Climate Change Studies
Genotype by Environment
Phenology

Environment



Step 1 : phenotyping data standardization and exchange



Heterogeneous and multidimensional

- Variability of experimental sites, designs, conditions
- Collected data - images, physical measurements, chemical assays, molecular biology assays, environmental,...

Dispersed

- No comprehensive public archive for phenotypic data

Poorly annotated

- Weaknesses in standards for data and metadata description

Plant data standards for FAIR

Semantic

- ◆ Description of the data
- ◆ Controlled vocabularies: term name and definitions
- ◆ Ontologies: semantic links between terms
- ◆ *Biologist driven*



Structure

- ◆ Formatting and Organizing the data
- ◆ Data Models
- ◆ Standards : CSV, VCF, GFF, MIAPPE (www.miappe.org) , etc...
- ◆ *Biologist & Computer scientist driven*



Technical

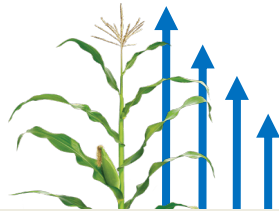
- ◆ Data integration and sharing
- ◆ Interoperability : tools and systems
 - ◆ GA4GH
 - ◆ Breeding API www.brapi.org
- ◆ *Computer scientist driven*



Phenotype Semantic Standard: Ontologies

- Describing traits/features in specific plant species
- Crop Ontology Trait + Method + Scale Semantic model

Variable identification: Plant height example



Trait

+

Method

+

Unit

M1: Total height

M2: First tassel branch

M3: Last expanded leaf

M4: Youngest growing leaf

...There is an uncountable number of combinations...

Each trait, method and unit has to be identified if we want to share and reuse data



T1: Plant Height

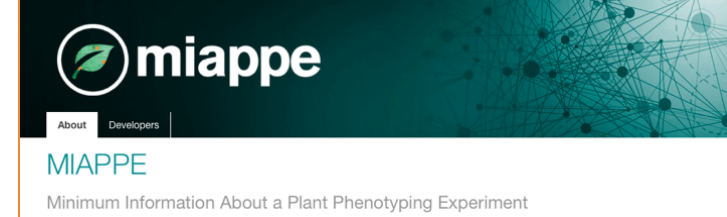


M5: Highest pixel
corresponding to plant

U3: pixel

Slide from L. Cabrera-Bosquet

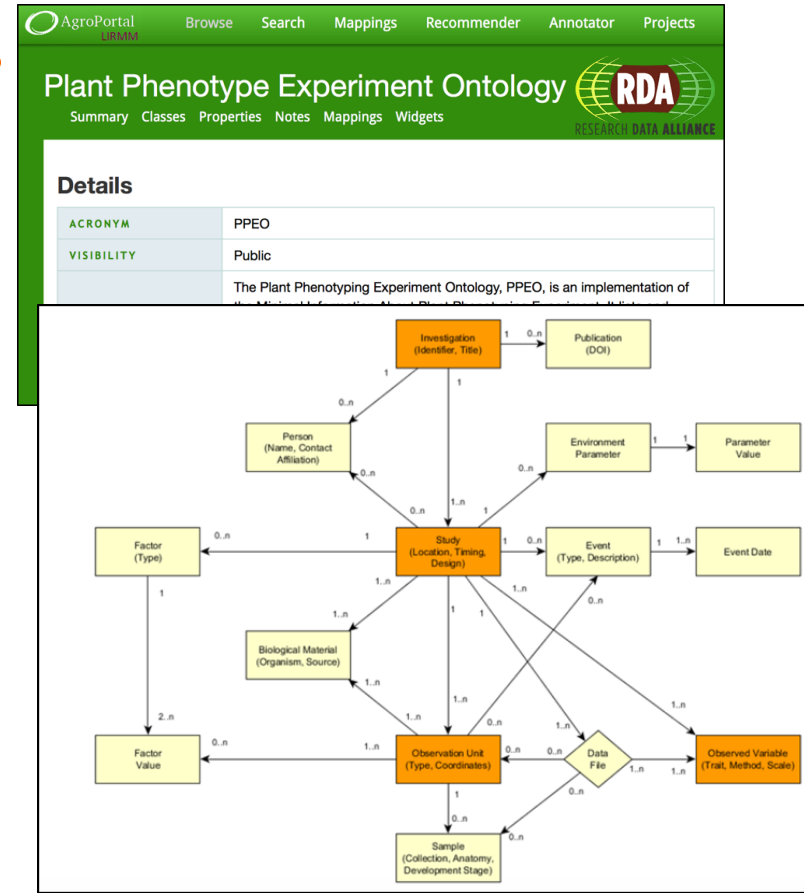
Phenotype Structure Standard



Minimal Information About Plant Phenotyping Experiment

www.miappe.org

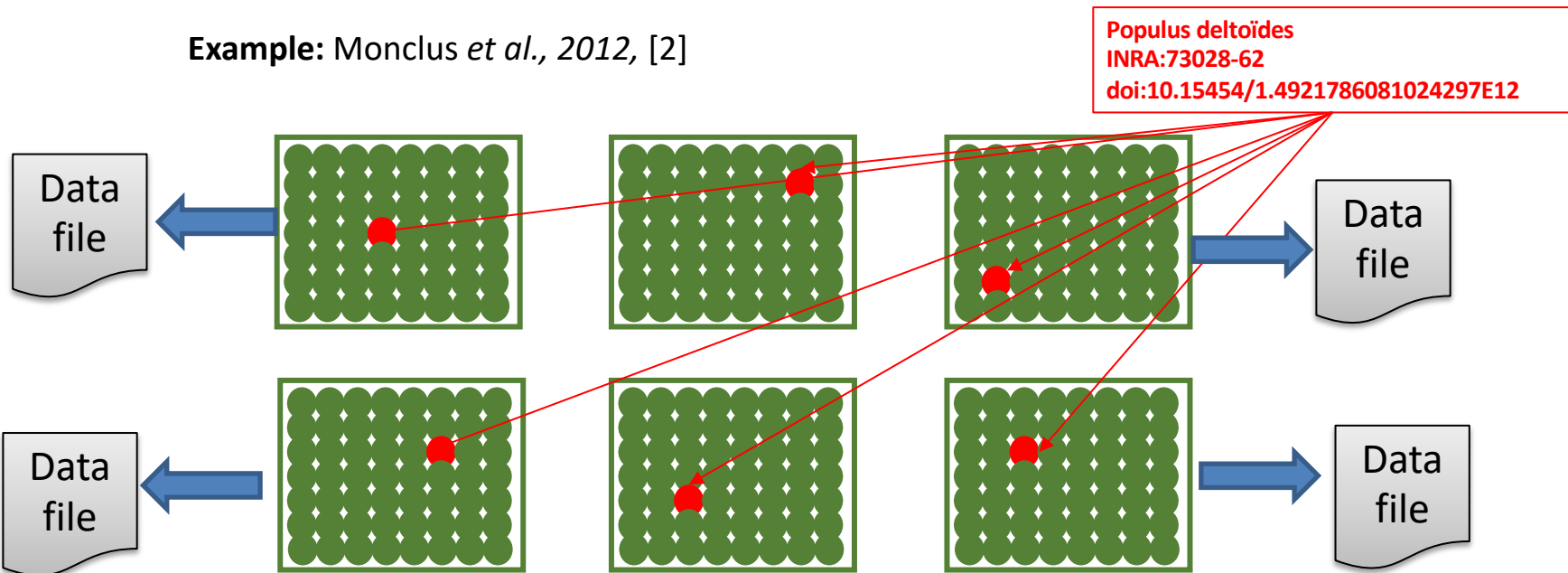
- Many stakeholders
- Open Community
- Crops and woody plants
- Biologist friendly
 - ◆ Clear definitions and examples
 - ◆ Excel templates
 - ◆ Trainings
- Computer scientist friendly
 - ◆ Explicit Data Model
 - ◆ ISA-Tools
 - ◆ Breeding API (BrAPI)
 - ◆ Phenotyping Experiment Ontology (PPEO) using OWL
(<http://agroportal.lirmm.fr/ontologies/PPEO>)



MIAPPE 1.1 Overview

Dataset

Example: Monclus *et al.*, 2012, [2]



6 randomized blocks

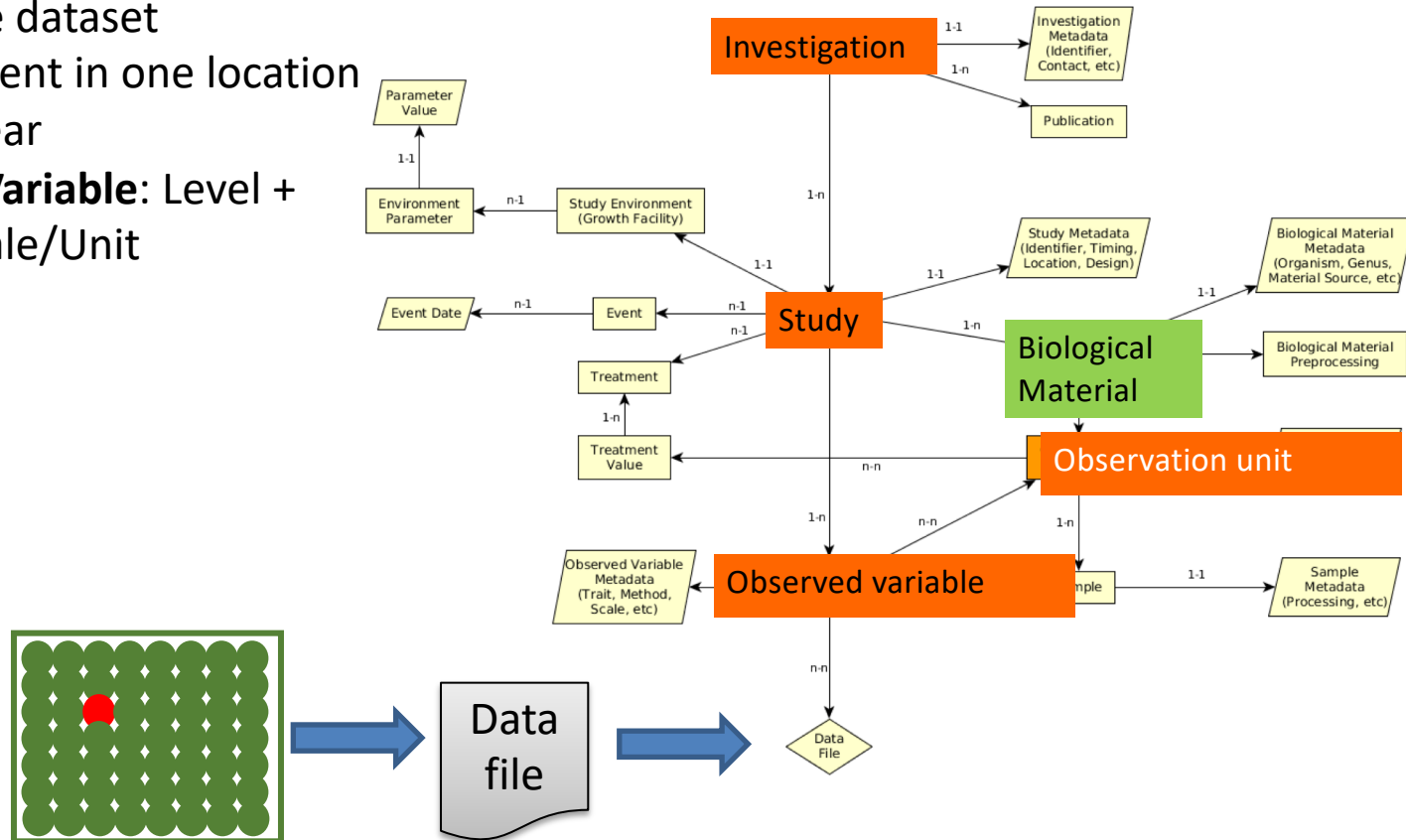
1 observation unit = one tree

No treatment

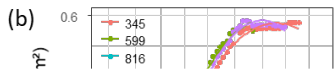
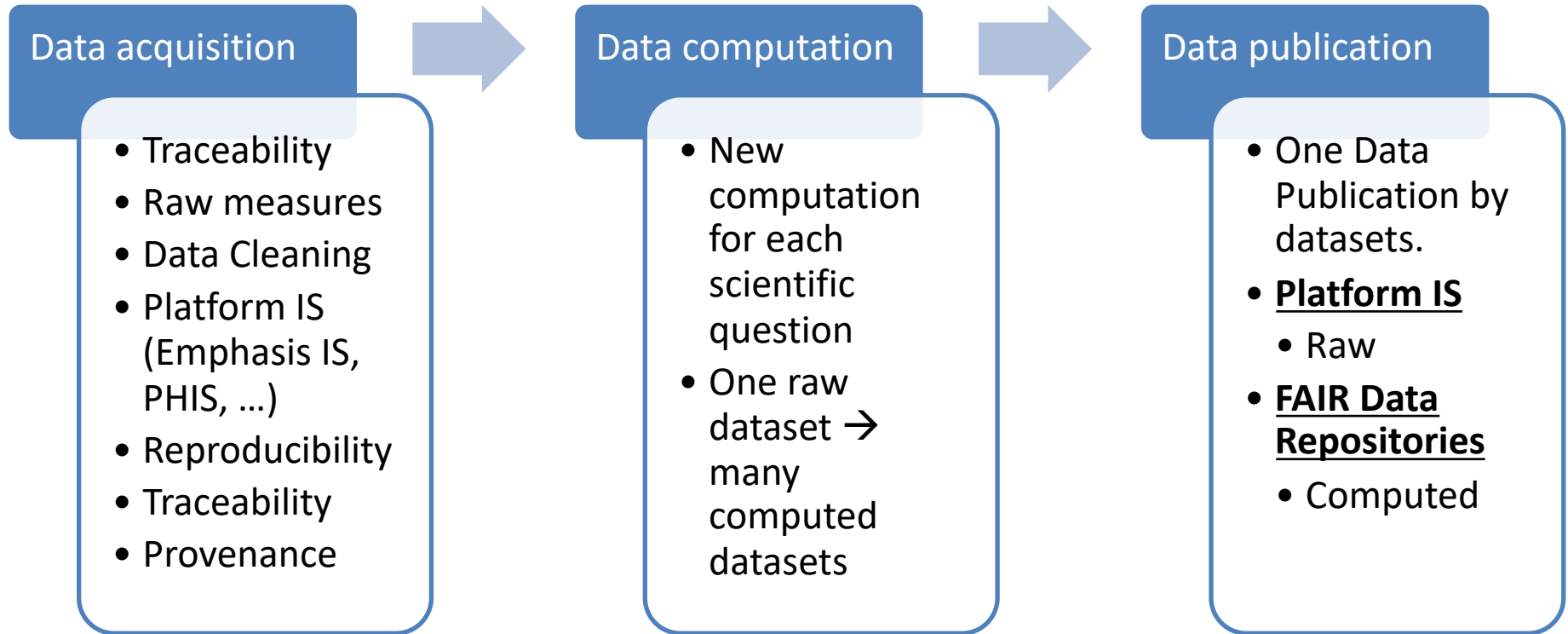
6 replicates defined by their position in each block: row and column

MIAPPE V1.1 Overview

- **Investigation:** whole dataset
- **Study** : one experiment in one location for one to several year
- **Assay *ie* Observed Variable:** Level + Trait + Method + Scale/Unit



Plant Phenotyping Life cycle



Data

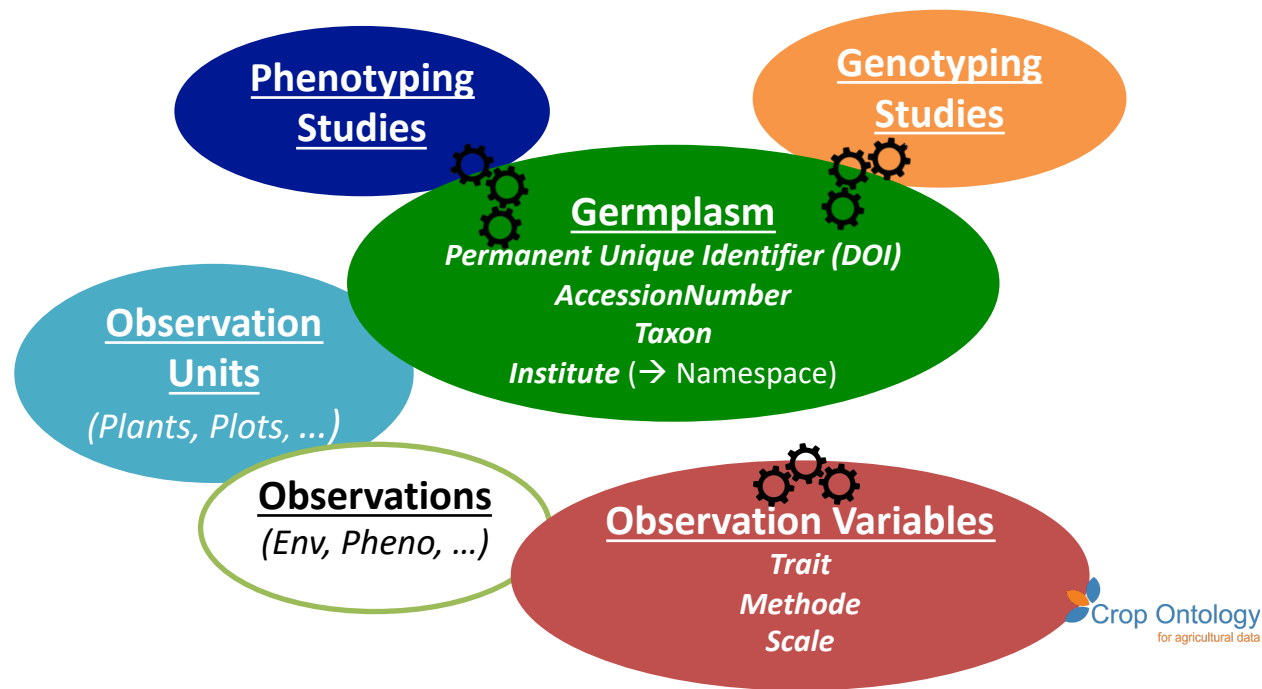
Genotype	traitement	Fusariose
Charger	low input	5
Charger	high N	2
Charger	low input	1
Charger	high N	2

Variety charger

Knowledge

Interventive cultural practice

Data Integration



Phenotype Technical Standard

- Breeding API <http://brapi.org/>
- International collaboration
 - ♦ Standard Open Web Service API, REST
 - ♦ Information Exchange, Main target: Breeding
 - ♦ Excellence in Breeding platform (CGIAR)
- Major INRA & Elixir Contribution
 - ♦ Phenotyping specification



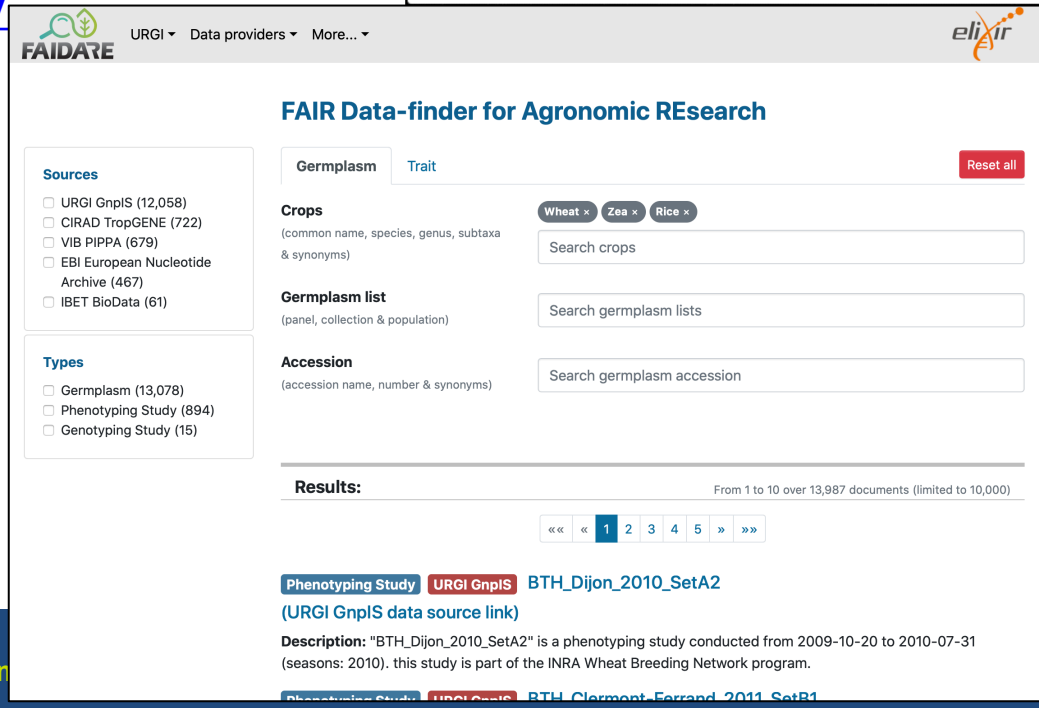
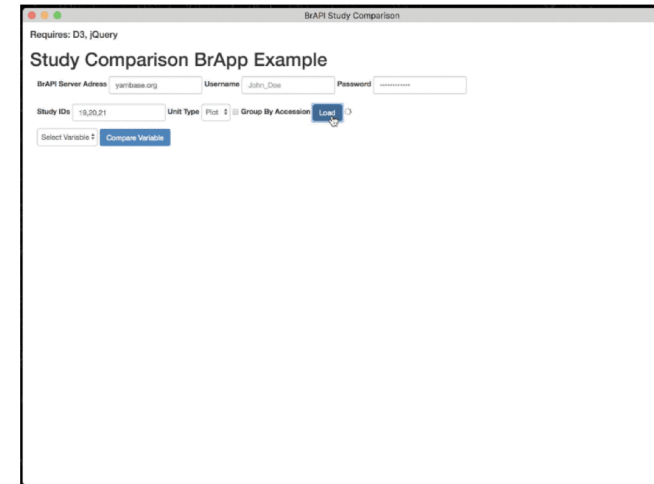
T3
CIRAD
GOBII
Wageningen
Cornell
iPlant

BILL & MELINDA
GATES foundation



Selby *et al. Bioinformatics* (2019),
doi.org/10.1093/bioinformatics/btz190

- Connect data repositories and tools:
 - ◆ Genotype visualization (Flapjack)
 - ◆ Studies graph preview and filtering
 - ◆ BrAPPS : Tools integrable in any BrAPI compliant System
 - ◆ <https://www.brapi.org/brapps.php>
 - ◆ R analysis environment
 - ◆ Field data capture
 - ◆ FAIR Data discovery → Elixir FAIDARE
 - ◆ <https://urgi.versailles.inra.fr/faidare/>



FAIDARE (FAIR Data-finder for Agronomic REsearch)

<https://urgi.versailles.inra.fr/faidare/>

- Open plant data repositories federation
- How to join us
 - ◆ Create your own BrAPI endpoint (documentation: <https://brapi.org/>)
 - ◆ Reference your BrAPI endpoint in FAIDARE (documentation: <https://urgi.versailles.inra.fr/faidare/join>)
- Information & Support
 - ◆ Elixir Plant Community : <https://www.elixir-europe.org/communities/plant-sciences>
 - ◆ INRA URGI: urgi-contact@inra.fr

The screenshot displays the FAIDARE web interface. At the top, there is a navigation bar with the FAIDARE logo, a dropdown menu showing 'URGI', 'Data providers', and 'More...', and the elixir logo. The main heading is 'FAIR Data-finder for Agronomic REsearch'. On the left, there are two filter panels: 'Sources' and 'Types'. The 'Sources' panel lists five options with checkboxes: URGI GnpIS (81,335), EBI European Nucleotide Archive (44,975), CIRAD TropGENE (722), VIB PIPPA (692), and IBET BioData (67). The 'Types' panel lists three options with checkboxes: Germplasm (94,589), Genotyping Study (32,210), and Phenotyping Study (992). The main content area has two tabs: 'Germplasm' (selected) and 'Trait'. Below the tabs, there are three search input fields: 'Search crops' (with a description: '(common name, species, genus, subtaxa & synonyms)'), 'Search germplasm lists' (with a description: '(panel, collection & population)'), and 'Search germplasm accession' (with a description: '(accession name, number & synonyms)'). A 'Reset all' button is located in the top right corner of the search area.

URGI ▾ Data providers ▾ More... ▾

FAIDARE

FAIR Data-finder for Agronomic REsearch

Sources

- ☐ URGI GnpIS (81,335)
- ☐ EBI European Nucleotide Archive (44,975)
- ☐ CIRAD TropGENE (722)
- ☐ VIB PIPPA (692)
- ☐ IBET BioData (67)

Types

- ☐ Germplasm (94,589)
- ☐ Genotyping Study (32,210)
- ☐ Phenotyping Study (992)

Germplasm Trait

Reset all

Crops
(common name, species, genus, subtaxa & synonyms)

Search crops

Germplasm list
(panel, collection & population)

Search germplasm lists

Accession
(accession name, number & synonyms)

Search germplasm accession

Create search criteria - Germplasm

Germplasm

Trait

Reset all

Crops
(common name, species, genus, subtaxa & synonyms)

Wheat ×

Search crops

Germplasm list
(panel, collection & population)

Search germplasm lists

Accession
(accession name, number & synonyms)

APACHE ×

war

Results:

WARRIOR

WARIMBA

WARDEN LEAP

38844

From 1 to 10 over 694 documents

Create search criteria - Trait

Germplasm **Trait** Reset all

Traits, methods and scales English NBF: Fruits/Cones number VARIABLE

Search terms...

☐ Vitis inra ontology ONTOLOGY

☐ Walnut Trait Ontology ONTOLOGY

☐ Wheat Inra Phenotype Ontology ONTOLOGY

☒ Woody Plant Ontology ONTOLOGY

☐ Abiotic stress TRAIT CLASS

☐ Agronomical TRAIT CLASS

☐ Biochemical TRAIT CLASS

☐ Biotic stress TRAIT CLASS

☐ Environmental TRAIT CLASS

☒ Fertility TRAIT CLASS

☒ Fructification TRAIT

☐ FRUC_AB: Fructification abundance VARIABLE

☒ NBF: Fruits/Cones number VARIABLE

☐ Fructification regularity TRAIT

Fructification TRAIT

Identifier CO_357:1000031

Name Fructification

Description Assessment of the fructification

Main abbreviation FRUC

Alternative abbreviations CON

Entity fruit

Attribute fructification

Status Standard for INRA & GenTree project

Class Fertility

Fruits/Cones protocol METHOD

Identifier CO_357:2000032

Name Fruits/Cones protocol

Description Number of mature fruits/cones that you are able to count on the tree in 30 seconds from the ground using binoculars

Reference GenTree_protocols_0.99.pdf pages 17

Class Counting

Count SCALE

Identifier CO_357:3000123

Traits, methods and scales English

canker

☐ Woody Plant Ontology ONTOLOGY

☐ Biotic stress TRAIT CLASS



☐ Canker girdling TRAIT

☐ Canker_GI: Canker girdling index VARIABLE

☐ Canker lesion length TRAIT

☐ Canker_LL: Canker lesion length VARIABLE

Find data

 URGI ▾ Data providers ▾ More... ▾ 

FAIR Data-finder for Agronomic REsearch

Sources

- ☐ EBI European Nucleotide Archive (3,134)
- ☐ CIRAD TropGENE (722)
- ☐ IBET BioData (61)

Types

- ☐ Germplasm (3,871)
- ☐ Phenotyping Study (28)
- ☐ Genotyping Study (18)

Germplasm Trait

Reset all

Crops
(common name, species, genus, subtaxa & synonyms)

Oryza ×
Search crops

Germplasm list
(panel, collection & population)

Search germplasm

Accession
(accession name, number & synonyms)

Search germplasm

Results:

From 1 to 10 over 3,917 documents

<< < 1 2 3 4 5 > >>

Genotyping Study

EBI European Nucleotide Archive

Oryza sativa cultivar:Geetanjali: IC523927 ; Sahabhadgadhan: IC560853 Raw sequence reads
(EBI European Nucleotide Archive data source link)

Description: "Oryza sativa cultivar:Geetanjali: IC523927 ; Sahabhadgadhan: IC560853 Raw sequence reads" is a Genotyping study. Rice is a chilling-sensitive plant and its exposure to low temperature during seedling stage results in slow seedling growth, yellowing, wither ...[\(Show More\)](#)

Phenotyping Study

IBET BioData

Rice phenotyping in salt stress
(IBET BioData data source link)

Description: "Rice phenotyping in salt stress" is a phenotyping study conducted from 2011-11-01 to 2011-11-21 (seasons: id, season, year).


Germplasm

CIRAD TropGENE


IR 36
(CIRAD TropGENE data source link)

Description: "IR 36" is a Oryza sativa L. (rice) accession.

Filter data



URGI ▾ Data providers ▾ More... ▾



FAIR Data-finder for Agronomic REsearch

Sources
☒ IBET BioData (2)

Types
☐ Germplasm (1)
☐ Phenotyping Study (1)

Germplasm Trait

Reset all

Crops
(common name, species, genus, subtaxa & synonyms)
Oryza ×
Search crops

Germplasm list
(panel, collection & population)
Search germplasm lists

Accession
(accession name, number & synonyms)
AMARELO ×
Search germplasm accession


Results:

From 1 to 10 over 2 documents

Germplasm **IBET BioData** **AMARELO**
(IBET BioData data source link)
Description: "AMARELO" is a Oryza sativa (Rice) accession (number: "IRGC9389").

Phenotyping Study **IBET BioData** **Rice phenotyping in salt stress**
(IBET BioData data source link)
Description: "Rice phenotyping in salt stress" is a phenotyping study conducted from 2011-11-01 to 2011-11-21 (seasons: id, season, year).

Data Access

Results:  From 1 to 10 over 3,917 documents

<< < 1 2 3 4 5 > >>

Genotyping Study **EBI European Nucleotide Archive** **Oryza sativa cultivar:Geetanjali: IC523927**

Sahabhadhan: IC560853 Raw sequence reads
(EBI European Nucleotide Archive data source link)



Description: "Oryza sativa cultivar:Geetanjali: IC523927 ; Sahabhadhan: IC560853 Raw sequence reads" is a Genotyping study. Rice is a chilling-sensitive plant and its exposure to low temperature during seedling stage results in slow seedling growth, yellowing, wither ...[\(Show More\)](#)

Phenotyping Study **IBET BioData** **Rice phenotyping in salt stress**
(IBET BioData data source link)


Description: "Rice phenotyping in salt stress" is a phenotyping study conducted from 2011-11-01 to 2011-11-21 (seasons: id, season, year).

Germplasm **CIRAD TropGENE** **IR 36**
(CIRAD TropGENE data source link)


Description: "IR 36" is a Oryza sativa L. (rice) accession.

 URGI Data providers More... 

Study : Oryza sativa cultivar:Geetanjali: IC523927 ; Sahabhadhan: IC560853 Raw sequence reads

Identification	
Name	Oryza sativa cultivar:Geetanjali: IC523927 ; Sahabhadhan: IC560853 Raw sequence reads
Identifier	dxJuoKwVQ59zdHVxS9QUkpQQT14ODg5Mg==
Source	
Data link	Link to this study on EBI European Nucleotide Archive
Description	Rice is a chilling-sensitive plant and its exposure to low temperature during seedling stage results in slow seedling growth, yellowing, withering, reduced tillering and stunting resulting in low productivity of high yielding but cold susceptible varieties. There are about 15 million hectares of rice fields in 24 different countries under threat from cold weather. Developments of high yielding varieties that are cold stress tolerant are likely to be rapidly adopted by the farmers in the target regions. The project aims to compare transcriptome of seedling stage cold tolerant variety with cold susceptible variety using RNA-seq approach during cold stress and recovery conditions to gain insight into different gene expression that might provide clues to the underlying principle of cold tolerance. In the study, Geetanjali has been identified as tolerant genotype for seedling stage cold stress at 40C, whereas susceptibility of Sahabhadhan has been confirmed. These two genotypes were used for transcriptome study.
Data files	SRR2089751 SRR2089753 SRR2089754 SRR2089756

Data access

Results:  From 1 to 10 over 3,917 documents

« « 1 2 3 4 5 » »

Genotyping Study **EBI European Nucleotide Archive** [Oryza sativa cultivar:Geetanjali: IC523927 ; Sahabhadghan: IC560853 Raw sequence reads](#)
(EBI European Nucleotide Archive data source link)



Description: "Oryza sativa cultivar:Geetanjali: IC523927 ; Sahabhadghan: IC560853 Raw sequence reads" is a Genotyping study. Rice is a chilling-sensitive plant and its exposure to low temperature during seedling stage results in slow seedling growth, yellowing, wither ...[\(Show More\)](#)

Phenotyping Study **IBET BioData** [Rice phenotyping in salt stress](#)
(IBET BioData data source link)

Description: "Rice phenotyping in salt stress" is a phenotyping study conducted from 2011-11-01 to 2011-11-21 (seasons: id, season, year).


Germplasm **CIRAD TropGENE** [IR 36](#)
(CIRAD TropGENE data source link)


Description: "IR 36" is a Oryza sativa L. (rice) accession.

 URGI Data providers More... 

Study : Oryza sativa cultivar:Geetanjali: IC523927 ; Sahabhadghan: IC560853 Raw sequence reads

Identification

Name	Oryza sativa cultivar:Geetanjali: IC523927 ; Sahabhadghan: IC560853 Raw sequence reads
Identifier	dxJuoKvWQ59zdHVXe59QUkpOQT140d95Mg==
Source	
Data link	Link to this study on EBI European Nucleotide Archive
Description	Rice is a chilling-sensitive plant and its exposure to low temperature during seedling stage results in slow seedling growth, yellowing, withering, reduced tillering and stunting resulting in low productivity of high yielding but cold susceptible varieties. There are about 15 million hectares of rice fields in 24 different countries under threat from cold weather. Developments of high yielding varieties that are cold stress tolerant are likely to be rapidly adopted by the farmers in the target regions. The project aims to compare transcriptome of seedling stage cold tolerant variety with cold susceptible variety using RNA-seq approach during cold stress and recovery conditions to gain insight into different gene expression that might provide clues to the underlying principle of cold tolerance. In the study, Geetanjali has been identified as tolerant genotype for seedling stage cold stress at 40C, whereas susceptibility of Sahabhadghan has been confirmed. These two genotypes

EMBL-EBI  **European Nucleotide Archive**

Home Search & Browse Submit & Update Software About ENA Support

The new ENA Browser is now live, with improved features for searching & downloading data! Please go to <https://www.ebi.ac.uk/ena/browser/view/PRJNA288892> to see the record there.

Study: PRJNA288892

Oryza sativa cultivar:Geetanjali: IC523927 ; Sahabhadghan: IC560853 Raw sequence reads

View: [Project XML](#) [Study XML](#) Download: [Project XML](#) [Study XML](#)

Name	Oryza sativa cultivar:Geetanjali: IC523927 ; Sahabhadghan: IC560853	Submitting Centre	Central Rice Research Institute	Organisation	Oryza sativa	Cultivar	Geetanjali: IC523927 ; Sahabhadghan: IC560853
Secondary accession(s)	GEOG00090						

Description

Rice is a chilling-sensitive plant and its exposure to low temperature during seedling stage results in slow seedling growth, yellowing, withering, reduced tillering and stunting resulting in low productivity of high yielding but cold susceptible varieties. There are about 15 million hectares of rice fields in 24 different countries under threat from cold weather. Developments of high yielding varieties that are cold stress tolerant are likely to be rapidly adopted by the farmers in the target regions.

The project aims to compare transcriptome of seedling stage cold tolerant variety with cold susceptible variety using RNA-seq approach during cold stress and recovery conditions to gain insight into different gene expression that might provide clues to the underlying principle of cold tolerance.

In the study, Geetanjali has been identified as tolerant genotype for seedling stage cold stress at 40C, whereas susceptibility of Sahabhadghan has been confirmed. These two genotypes were used for transcriptome study.

Lineage

Eukaryota, Viridiplantae, Streptophyta, Embryophyta, Tracheophyta, Spermatophyta, Magnoliopsida, Liliopsida, Poales, Poaceae, BOP clade, Oryzoideae, Oryzoaceae, Oryzinae, Oryza

Navigation [Read Files](#) [Portal](#) [Attributes](#) [Parent Projects](#)

[Bulk Download Files](#)

FAIDARE web services

<https://urgi.versailles.inra.fr/faidare/swagger-ui.html>

The screenshot shows the Swagger UI for the FAIDARE API. The header is green with the 'swagger' logo and a 'Select a spec' dropdown set to 'default'. The main content area is titled 'Api Documentation' with a version '1.0' badge. Below the title, it lists the base URL as 'urgi.versailles.inra.fr/faidare' and provides a link to the API docs. There are links for 'Api Documentation', 'Terms of service', and 'Apache 2.0'. The API is categorized into 'Breeding API' (BrAPI endpoint) and 'FAIDARE API' (Extended FAIDARE API). The 'FAIDARE API' section is expanded, showing a list of endpoints: a POST endpoint for searching data discovery documents, a GET endpoint for listing data sources, a POST endpoint for suggesting document field values, a GET endpoint for searching germplasm by ID or PUI, a GET endpoint for exporting germplasm as CSV, and a GET endpoint for finding xref documents by full text ID. At the bottom, there is a 'Models' section.

swagger Select a spec default

Api Documentation ^{1.0}

[Base URL: urgi.versailles.inra.fr/faidare]
<https://urgi.versailles.inra.fr/faidare/v2/api-docs>

Api Documentation
[Terms of service](#)
[Apache 2.0](#)

Breeding API BrAPI endpoint >

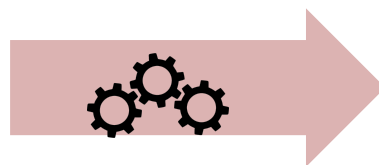
FAIDARE API Extended FAIDARE API ✓

- POST** `/faidare/v1/datadiscovery/search` Search for data discovery documents
- GET** `/faidare/v1/datadiscovery/sources` Get list of data sources
- POST** `/faidare/v1/datadiscovery/suggest` Suggest data discovery document field values
- GET** `/faidare/v1/germplasm` Search germplasm by ID or PUI
- GET** `/faidare/v1/germplasm/csv` export
- GET** `/faidare/v1/xref/documentbyfulltextid` Find xref documents

Models >

IA & New analysis tech for Data Integration

Variety charger is resistant to fusariose under intensiv cultural practice



Charger
Triticum aestivum

Resistant under
intensiv cultural
practice

Fusariose
Fusariose
NA
NA

Phenotyping
Studies

Genotyping
Studies

Germplasm

Permanent Unique Identifier (DOI)
AccessionNumber
Taxon
Institute (→ Namespace)

Observation
Units

(Plants, Plots)

Observations
(Env, Pheno,
...)

Observation Variables

Trait
Methode
Scale

M2 Master Internship on Text mining

- URGI / Bibliome partnership
- <https://urgi.versailles.inra.fr/About-us/News/Text-mining-data-integration-web-devel-M2-internship>

The screenshot shows the URGI website interface. At the top, there's a navigation bar with links like 'FEEDBACK', 'CONTACT', 'SITE MAP', and 'ABOUT US'. The main header features the URGI logo and a 'PLANT A' banner. Below the header, there's a green navigation bar with 'Platform', 'Research', and 'Projects'. The left sidebar contains 'About us', 'Publications', and 'News' (with a year list from 2019 to 2011). The main content area displays a large 'WE WANT YOU!' graphic with a fist icon. To the right, there's a 'Register' button and a search bar. The text 'Integration / web devel M2 internship' is visible, along with a '2 internship' label. The bottom of the page features a blue banner with the INRA logo and the text 'Cyril Pommier / Heterogenous plant data from data integration practices to analysis perspectives.'

WE WANT YOU!

04 Oct 20

A 6 month position is available at URGI for a Master 2 internship.

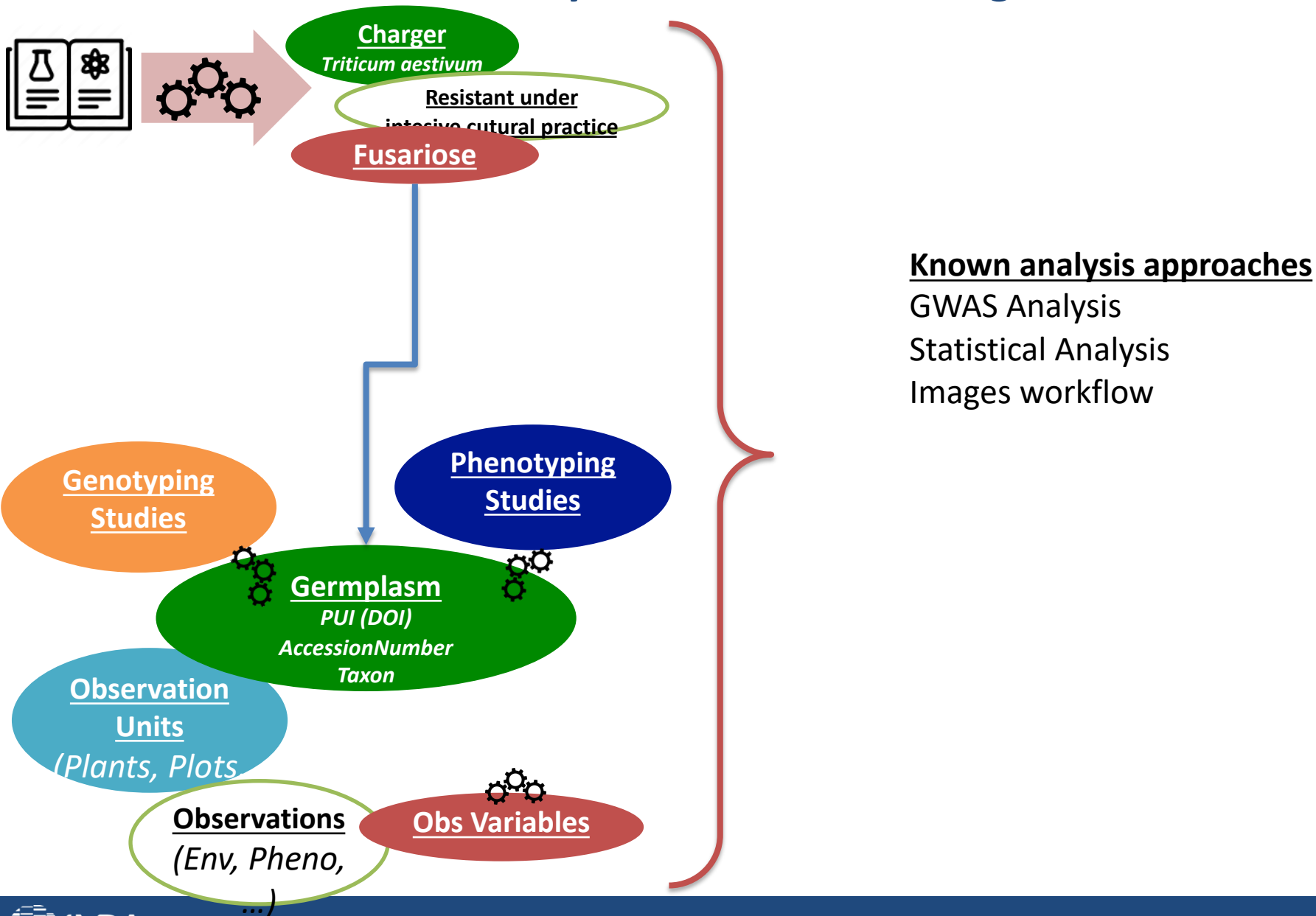
Applicant will perform integration of wheat text-mining data, will automate and generalize this process to other species and will develop web interfaces to make the data publicly available.

Application available until 2019/12/31.

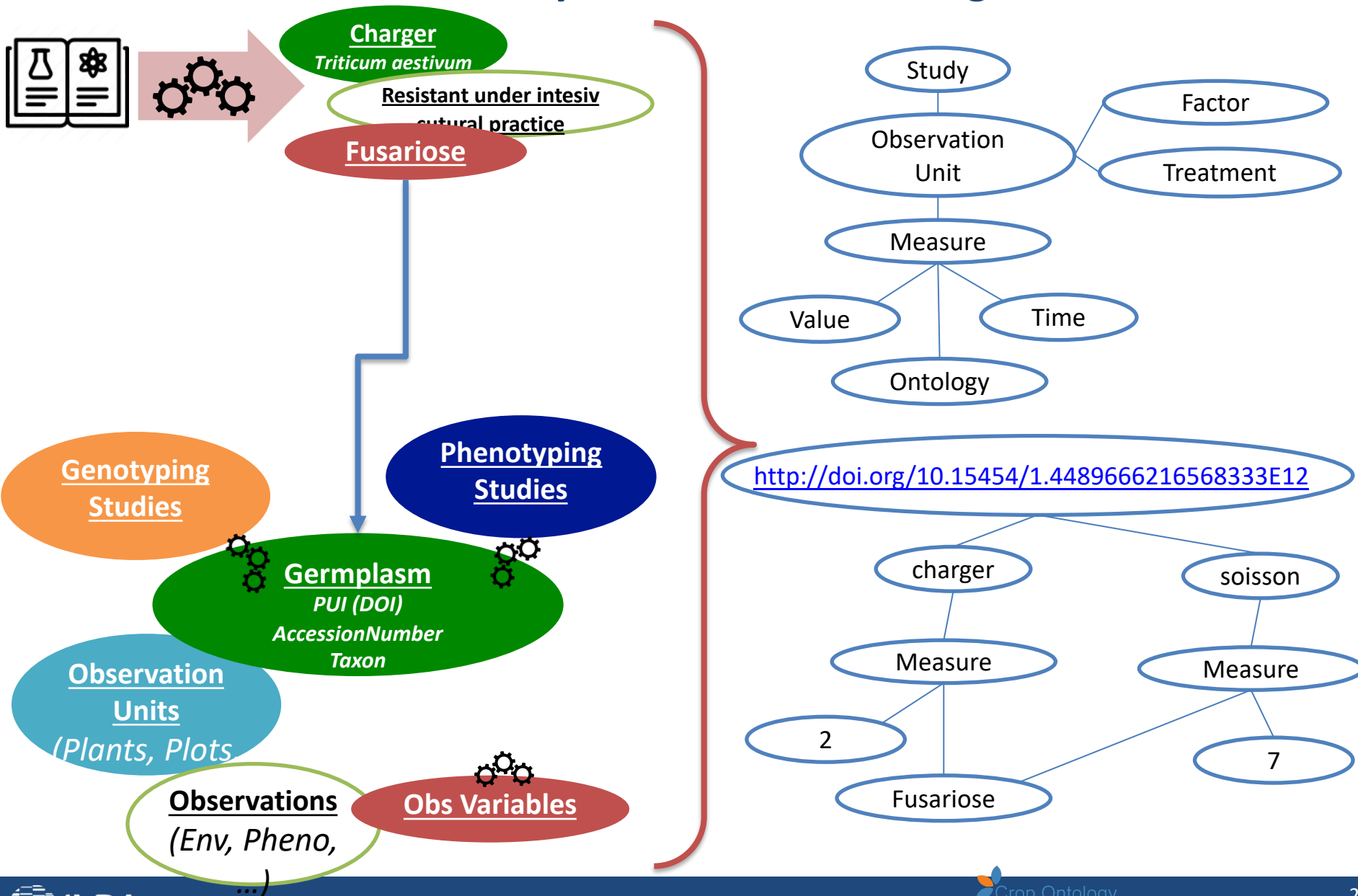
More details available in french in the following file:
[2019-08-M2_Bioinfo_URGI_TextMine_fr.pdf](#) 162.39 kB

2 internship

IA & New analysis tech for Data Integration



IA & New analysis tech for Data Integration



MIAPPE Semantic: RDA RDFENO

- MIAPPE OWL Implementation
 - ◆ First draft:
<https://github.com/MIAPPE/MIAPPE-ontology>
- Brapi 2 MIAPPE RDF workflow
 - ◆ <http://ist.blogs.inra.fr/wdi/phenotypes-as-rdf/> (www.wheatIS.org data standards)
 - ◆ JSON LD based
- Publication on MIAPPE.ORG
- Wheat dataset
 - ◆ <http://dx.doi.org/10.15454/1.4489666216568333E12>

Wheat Data Interoperability Guidelines

Home Guidelines Ontologies & Vocabularies Use cases Getting involved About

Home / Phenotypes as RDF

Phenotypes as RDF

Phenotypes to Semantic web publication

The sharing and especially the reusability of complex data like phenotypes is a good use case for a publication following the semantic web principles. This page describes how to generate a Phenotype RDF dataset from a [Breeding API](#) MIAPPE compliant endpoint. The growing adoption of this API will facilitate the reusability of such a transformation process.

Recommendations

The RDF publication relies among other things on the Permanent Unique Identifier (PUI), either URIs or DOIs

Virtuoso SPARQL Query Editor

Default Data Set Name (Graph IRI)

Query Text

```
PREFIX brapi: <https://brapi.org/rdf/>
SELECT ?network ?study ?attribute ?value
FROM <urn:urci:pheno-brapi-inra-small-grain-cereals-network>
WHERE {?study a brapi:Study;
          ?attribute ?value.
        ?trial a brapi:Trial;
              brapi:hasName ?network.}
ORDER BY ?network ?study
```

(Security restrictions of this server do not allow you to retrieve remote RDF data, see [details](#).)

Results Format:

HTML

Execution timeout:

0 milliseconds (values less than 1

Options:

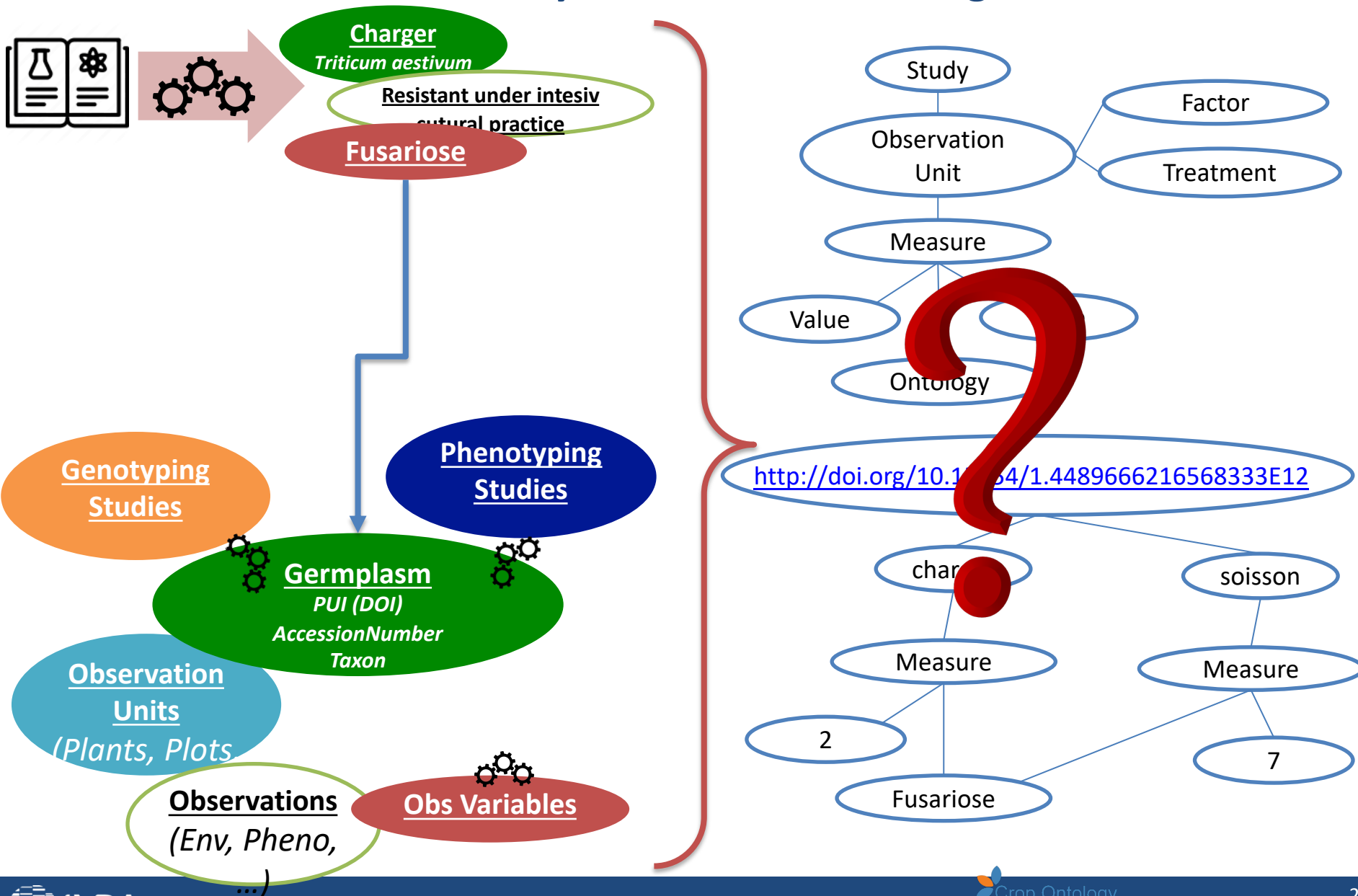
☒ Strict checking of void variables ☐ Log debug info

(The result can only be sent back to browser, not saved on the server, see [details](#))

Run Query

Reset

IA & New analysis tech for Data Integration



THANK YOU

