



**HAL**  
open science

## FAIR Data publication for phenomics data

Cyril Pommier

► **To cite this version:**

Cyril Pommier. FAIR Data publication for phenomics data. Integrative Bioinformatics Symposium 2018, BBSRC Rothamsted Research. GBR., Jun 2018, Rothamsted, United Kingdom. pp.52 slides + 40 slides. hal-02787988

**HAL Id: hal-02787988**

**<https://hal.inrae.fr/hal-02787988v1>**

Submitted on 5 Jun 2020

**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.



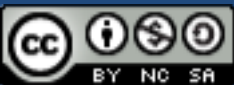
Distributed under a Creative Commons Attribution - NonCommercial - ShareAlike 4.0 International License



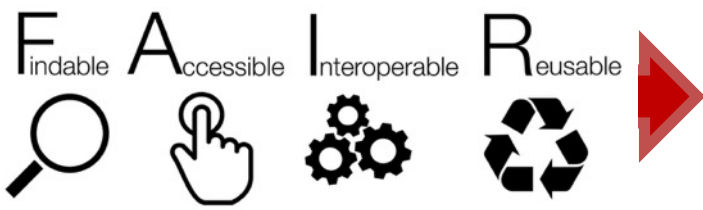
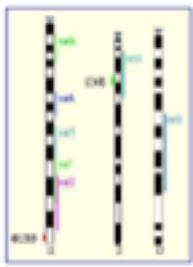
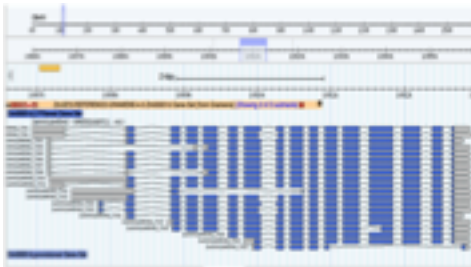
# Standards for Distributed Plant data integration



## Challenges and Solutions



# Genomic X Environment X Phenomic



- Climate change adaptation
- Variety improvement
- Cultural Practices adaptations
- ...



# FAIR Data Principles

Ids  
Metadata  
Index

Open Protocols  
Persistent Metadata

Semantics  
Linked Data  
Vocabularies

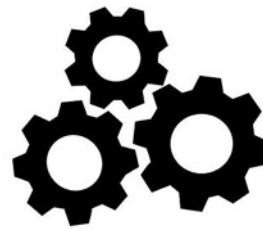
License  
Well described  
Provenance  
Standards

F  
Findable

A  
Accessible

I  
Interoperable

R  
Reusable



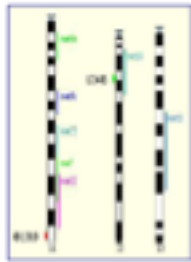
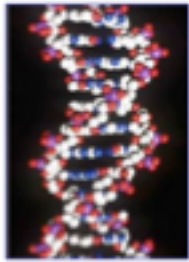
- Avoid Data Silos
- Interoperable Data and Tools Integration
  - ◆ Linked data
- Repositories Complementarity
  - ◆ Volume and Variety Burden Sharing

# FAIR for Genomic

- High Throughput
- Genome assemblies
- Genetic variations
- Established repositories
  - ◆ ENA/EVA
  - ◆ NCBI
  - ◆ Consortium/Species repositories
- Established standards
- Simple life cycle
  - ◆ Production → one data publication



# FAIR for Phenomic $P = G \times E \times C$

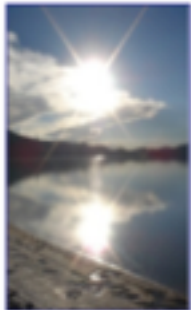


**Plant Genotype**

**Interactions**



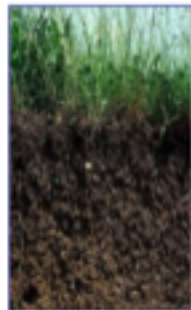
**Environment**



Climate



Cultural Practice



Soil



Stress

**Phenotype**



Phenology, Structure, Biochemical content

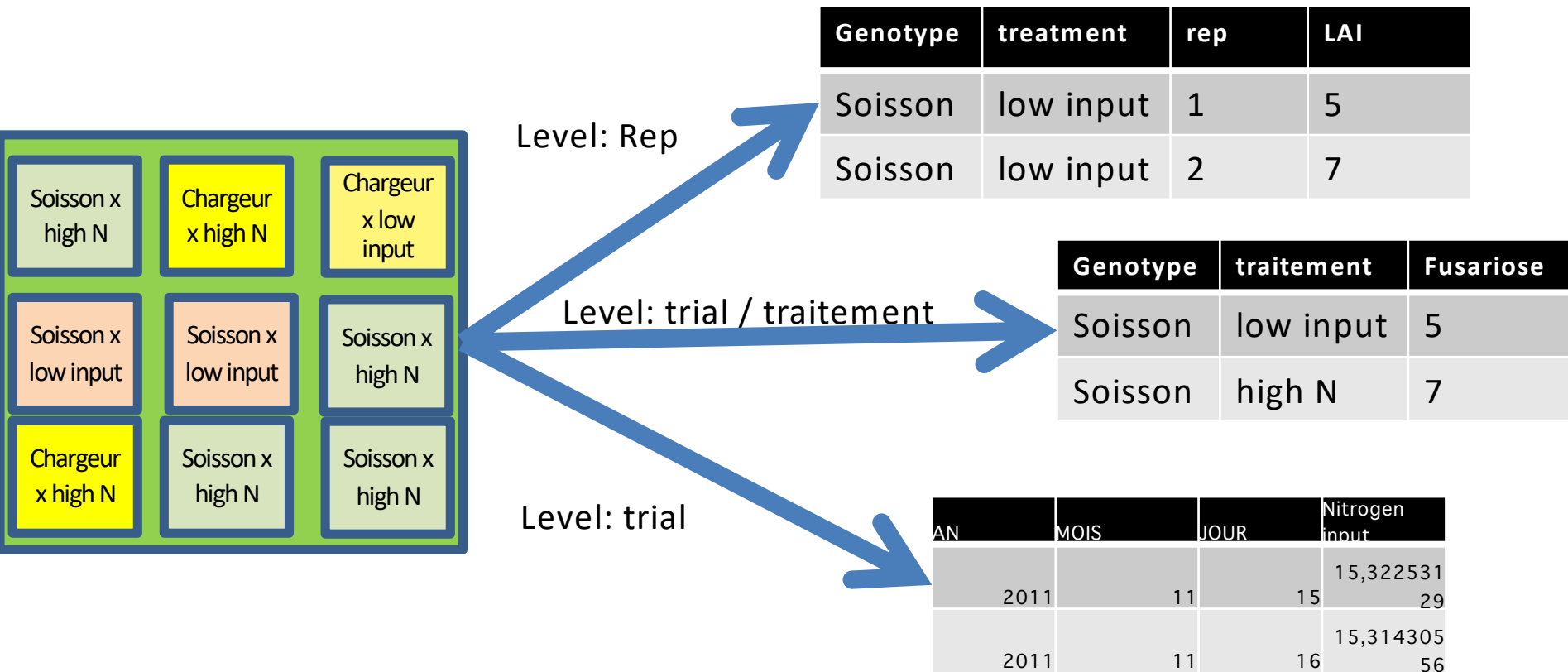
- Trait/Variable : a definition of an observable characteristic
  - ◆ Yield
- Phenotype : A trait/Variable with a value
  - ◆ Yield = 15 Kg by hect

# FAIR for Phenomic

- P = Genotype x Environment x Cultural practice
- Environment is critical, included in Phenotype lifecycle
- Raw data
  - ◆ Big : Terabyte by experiment
  - ◆ Patrimonial: long term storage
- No central Phenomic data archive
- Distributed Model
  - ◆ Wheat IS, Elixir, Transplant
  - ◆ Standards : Breeding API, MIAPPE, Crop Ontology
  - ◆ Several repositories: GnpIS, eDale, Emphasis (PHIS), PIPA/VIB, ...
- Complex life cycle

# Phenotyping Data life cycle

## Raw data

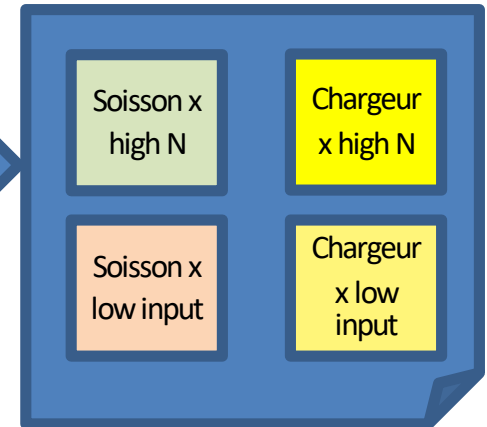
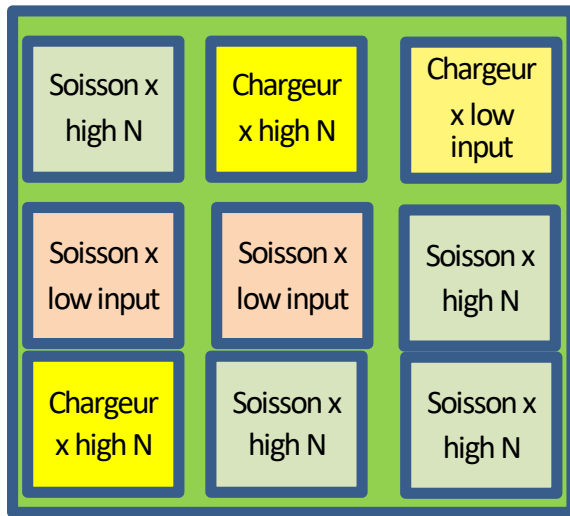




# Phenotyping Data Computation

Raw data

Elaborated data



Genotype	treatment	rep	LAI
Soisson	fi	1	5
Soisson	fi	2	7

Genotype	treatment	LAI
Soisson	low input	6

# Phenotyping Data Computation

- NIRS

Individual	Genotype	Repetition	Date	4000	4002	4004	4006
662200657-1	662200657	1	Nov11	0.790383794657 181	0.790834341907 288	0.791261293580 598	0.791659233526 804
662200657-2	662200657	2	Nov11	0.817254078896 638	0.817865344054 208	0.818450087726 549	0.819007314250 037
662200659-1	662200659	1	Nov11	0.822434379133 726	0.822871701993 426	0.823268552780 517	0.823649256266 632

- After Calibration and Analysis

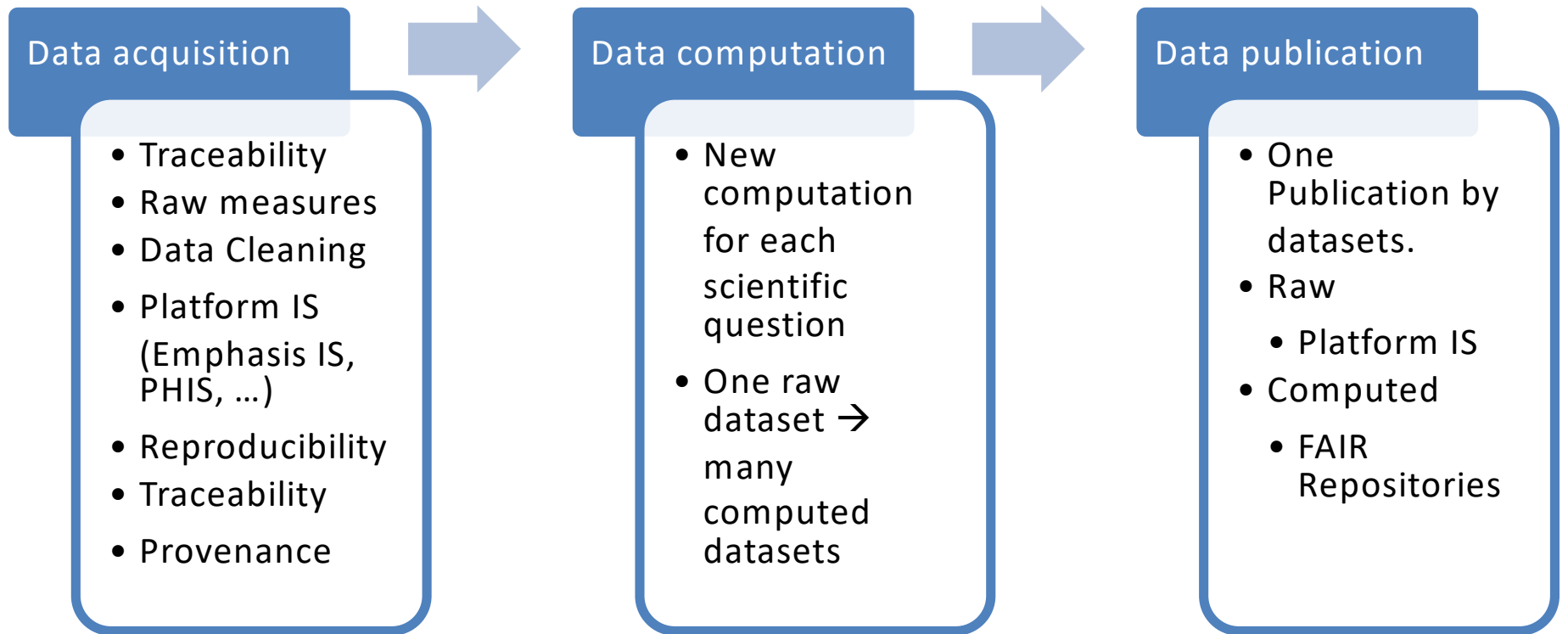
Individual	Genotype	Repetition	Gluc_sol	Gluc_solub	Gluc_hydrol	NIRS_id	Date
662200659-1	662200659	1	0.064074733	0.38557928	0.321504547	2274	Nov11

# Phenotyping Data Computation

- LIDAR
  - ◆ Very Big files
  - ◆ Plant Architecture
  - ◆ Raw extraction :
    - Leaf area, leaf angle
  - ◆ Elaboration → % interception
- Metabolomic
  - ◆ MSI Compliant Full description
  - ◆ Elaboration →
    - Leaf Protein concentration
    - Root Cellulose concentration

# Plant Phenotyping Life cycle

## Raw data long term conservation



IB 2018 Workshop  
main Topic

# Other type Lifecycle

- Heterogenous data types
  - ◆ Markers
  - ◆ Genetic resources accessions
  - ◆ Genetic experiment results
    - GWAS
    - Genomic Selection
    - Genetic Maps
  - ◆ Journal articles
  - ◆ Seeds of discovery data
- Minimal FAIR
  - ◆ Simple life cycle : publication
  - ◆ Findability : datadiscovery.

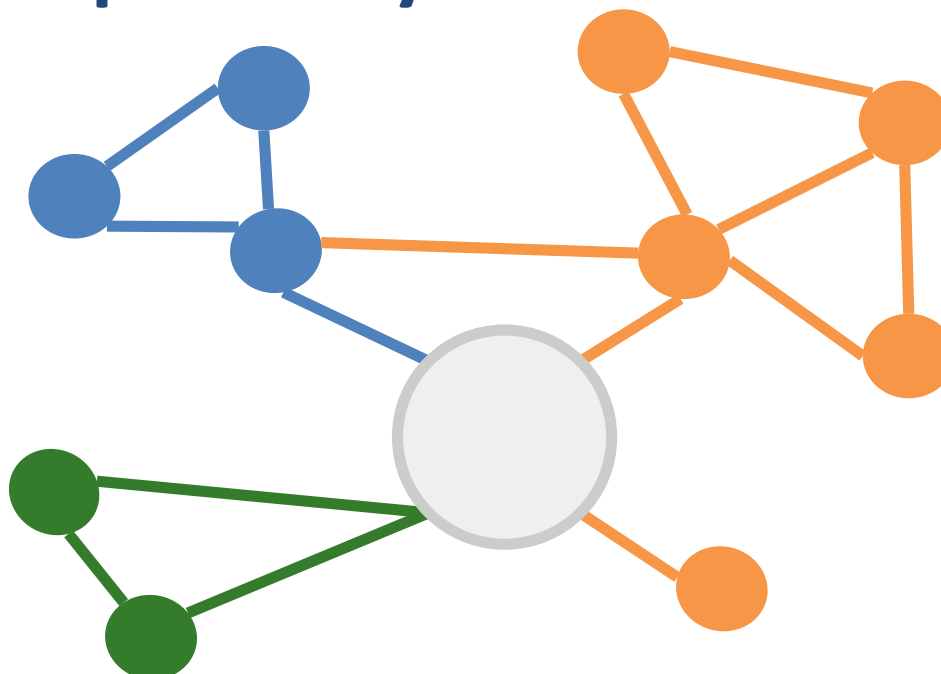
## FAIR: DATA STANDARDS



# Interoperability in International network

**National Networks**

**European Networks**



**Global Networks**

**International data standards**



Web services



Minimal information



Controlled vocabularies  
Trait dictionaries



MCPD



# Standard Sharing

How to cite this record FAIRsharing.org: GnplS; Genetic and Genomic Information System; DOI: 10.25504/FAIRsharing.dw22y3; Last edited: May 8, 2018, 9:00 a.m.; Last accessed: Jun 12 2018 9:23 p.m.

This record is maintained by [cpommier](#) [ORCID](#) and [ThomasLetellier](#)

Record added: March 2, 2016, 5:59 a.m.

Record updated: May 7, 2018, 11:29 a.m. by [ThomasLetellier](#).

## In Collections

[Wheat Data Interoperability Guidelines](#)

[ELIXIR node contributed resources](#)

## Support

[General](#)

## Related Standards

### Reporting Guidelines

[Minimum Information about Plant Phenotyping Experiment](#)

### Terminology Artifacts

[Crop Ontology](#)

[Plant Ontology](#)

### Models and Formats

[Investigation Study Assay Tabular](#)

[Generic Feature Format Version 3](#)

[Variant Call Format](#)



# WheatIS standards

- <http://wheatis.org/DataStandards.php>
- Phenotypes guidelines
- MIAPPE
- Croponontology
- Breeding API

## Wheat Data Interoperability Guidelines

Home Guidelines **▼** Ontologies & Vocabularies Use cases **▼** Getting involved About **▼**

Home / Phenotypes

### Phenotypes

Characteristics of an organism resulting from interactions between genes and the environment. Plant phenotyping can be used in several fields, such as breeding and crop improvement. Phenotypes (traits and associated value) can be used in several fields of research in multi-location environments, germplasm bank and crop improvement. Furthermore, phenotypes are in interaction, like for instance in crop improvement.

Getting Involved



## 1. Data formats

We recommend following minimum format principles with data variables (trait along with method, units and scales or environmental conditions). This principle has been formalized in the **MIAPPE** under improvement.

The **Breeding API**, described below, is a MIAPPE compliant web service API specification.

There is also an emerging initiative to produce a semantic web version of MIAPPE through a **to RDF transformation (new!)**.

See the **germplasm recommendations** for data format regarding germplasm information.

# WheatIS standards

- Reuse for other crops: Rice, ...

## Wheat Data Interoperability Guidelines

[Home](#) [Guidelines](#) [Ontologies & Vocabularies](#) [Use cases](#) [Getting involved](#) [About](#)

[Home](#) / [Getting involved](#)

### Getting involved

## See who endorses the guidelines

To stay relevant and useful for the wheat data community, it is important that the recommendations contained here are maintained. This will be especially true as new technologies are developed requiring standards to be agreed. The maintenance of the guidelines will depend on an ongoing process of monitoring practices across the wheat community, and maintaining communication with those who are producing and using wheat data. There are many ways you can help.

**Become an adopter:** Please let us know why and how these guidelines are helpful to you and your institution. See the [Current list of adopters](#)

**Give your feedback:** You can contact us either by email or by leaving a comment on a specific page. This is an easy way to propose improvements or corrections, update the state of the art and share best practices and useful tools.

### Getting Involved



Whe

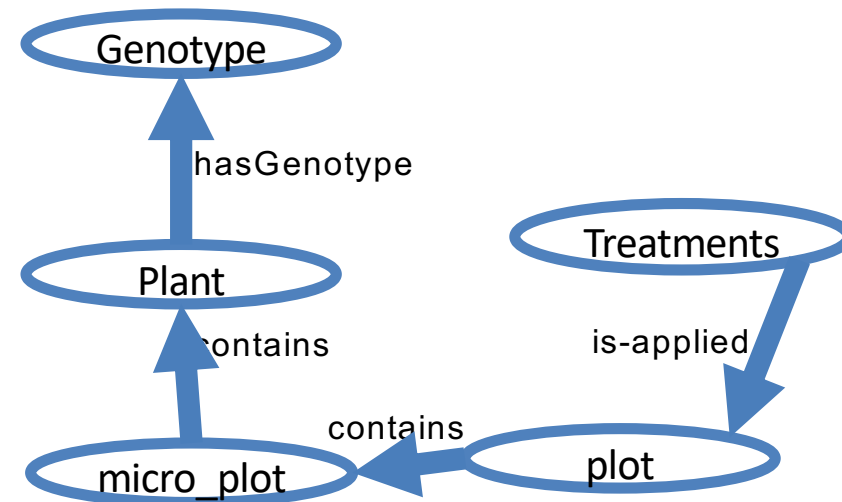
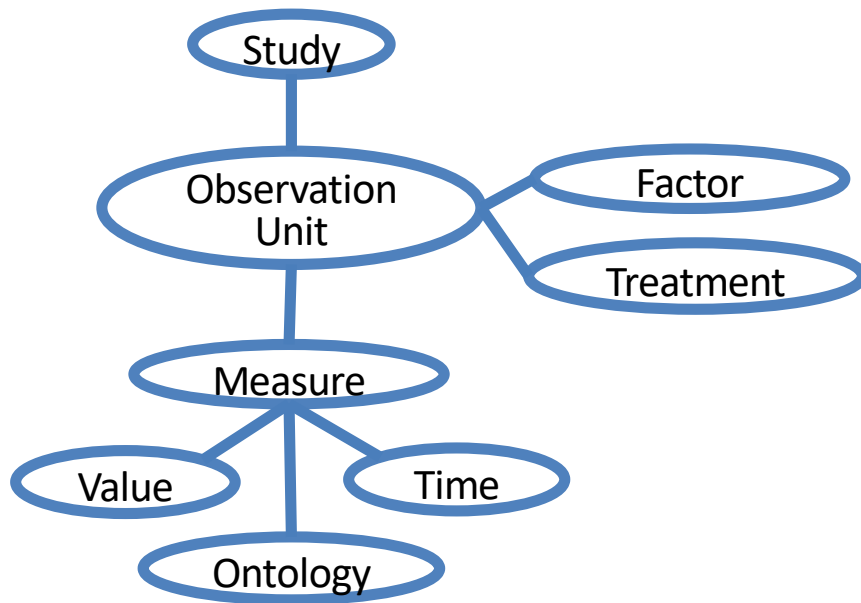
# INTEROPERABILITY: SEMANTIC



FAIR

# Semantic

- Define concepts: plant, micro\_plot, phenotype, experiment, event
- Define Relations: micro\_plot contains plant
- Common Vocabularies and ontologies
  - ◆ Agrovoc, Plant Ontology, Crop Ontology, Environment Ontology, ...



Elizabeth Arnaud et al.

## Phenotype data in breeding

Annotation must explain:

Germplasm ID	PH	PH2	GCOL	GY
24530	80	1	2	35
85432	120	3	4	48
78452	95	2	4	43
56093	100	2	1	50

1/ What is the observation about ?

=

**TRAIT or Environment**

The height of the plant

The color of the grain

The grain yield

Elizabeth Arnaud et al.

## Phenotype data in breeding

Annotation must explain:

Germplasm ID	PH	PH2	GCOL	GY
24530	80	1	2	35
85432	120	3	4	48
78452	95	2		43
56093	100	2	1	50

2/ How is the trait observed ?

=

**METHOD**

With a measuring tape

Estimated visually

Calculated  
Panicle weight x panicle density

Elizabeth Arnaud et al.

## Phenotype data in breeding

Annotation must explain:

Germplasm ID	PH	PH2	GCOL	GY
24530	1	1	2	35
85432	120	3	3	48
78452	95	1	3	43
56093	100	2	1	50

3/ How is the trait observation expressed?

=

**SCALE**

cm

1 = short (<90cm)  
2 = medium  
(>90cm and <110cm)  
3 = tall (>110cm)

1 = white  
2 = cream  
3 = yellow

g/m<sup>2</sup>

Elizabeth Arnaud et al.

## Phenotype data in breeding

Germplasm ID	PH	PH2	GCOL	GY
24530	80	1	2	35
85432	120	3	3	48
78452	95	2	3	43
56093	100	2	1	50

**1 VARIABLE**  
=  
**1 TRAIT/ENV**  
**+ 1 METHOD**  
**+ 1 SCALE**

**Woody Plant Ontology** ONTOLOGY

Ontology name: Woody Plant Ontology  
 Authors: Célia Michotey  
 Version: v1  
 Licence: CC BY 4.0  
 Links: [CropOntology.TVD5 format](#)  
[Adonis format](#)

**Traits, methods and scales**

Search terms...

Woody Plant Ontology ONTOLOGY

- Agronomical TRAIT CLASS
- Fertility TRAIT CLASS
- Morphological TRAIT CLASS
- Other TRAIT CLASS
- Phenological TRAIT CLASS
- Physiological TRAIT CLASS
- Quality TRAIT CLASS
- Stress TRAIT CLASS
- Stress: Abiotic stress TRAIT CLASS
- Stress: Biotic stress TRAIT CLASS
  - Bacterial canker girdling index TRAIT
  - Bacterial canker lesion length TRAIT
  - Game damages TRAIT
  - Nests presence TRAIT
  - NID: Nests presence VARIABLE

**NID: Nests presence** VARIABLE

Identifier: WOODY:0000067  
 Name: NID  
 Synonyms: Nests presence  
 NID(Adonis)  
 Context of use: Trial evaluation  
 Status: Standard for INRA  
 Institution: INRA  
 Scientist: Célia Michotey  
 Date: 13/03/2017  
 Crop: Woody Plant

**Nests presence** TRAIT

Identifier: WOODY:1000056  
 Name: Nests presence  
 Description: Presence or absence of nests  
 Main abbreviation: NID  
 Status: Standard for INRA  
 Class: Stress: Biotic stress

**Visual scoring** METHOD

Identifier: WOODY:2000003  
 Name: Visual scoring  
 Description: Visual assessment with a reference scoring scale  
 Class: Estimation

**Presence/ absence** SCALE

Identifier: WOODY:3000003  
 Name: Presence/Absence  
 Data type: Nominal  
 Decimal places: 0  
 Min: 0



# INTEROPERABILITY: STANDARD



# Minimum Information About Plant Phenotyping Experiment

- Standard
  - ◆ Data exchange and traceability
  - ◆ Repositories
- Input and output for analysis pipelines developed for phenotyping and genetic
- Implementations
  - ◆ Repositories/Databases
  - ◆ Web Service 
  - ◆ RDF
    - ◆ <http://purl.org/pppeo>

## Measures for interoperability of phenotypic data: minimum information requirements and formatting

Hanna Ćwiek-Kupczyńska, Thomas Altmann, Daniel Arend, Elizabeth Arnaud, Dijun Chen, Guillaume Cornut, Fabio Fiorani, Wojciech Frohmberg, Astrid Junker, Christian Klukas, Matthias Lange, Cezary Mazurek, Anahita Nafissi, Pascal Neveu, Jan van Oeveren, Cyril Pommier, Hendrik Poorter, Philippe Rocca-Serra, Susanna-Assunta Sansone, Uwe Scholz, Marco van Schriek, Ümit Seren, Björn Usadel, Stephan Weise, Paul Kersey and Paweł Krajewski 

*Plant Methods* 2016 12:44 | DOI: 10.1186/s13007-016-0144-4 | © The Author(s) 2016

Received: 15 April 2016 | Accepted: 18 October 2016 | Published: 9 November 2016

## Towards recommendations for metadata and data handling in plant phenotyping

Paweł Krajewski<sup>1,\*</sup>, Dijun Chen<sup>2</sup>, Hanna Ćwiek<sup>1</sup>, Aalt D.J. van Dijk<sup>3</sup>, Fabio Fiorani<sup>4</sup>, Paul Kersey<sup>5</sup>, Christian Klukas<sup>2</sup>, Matthias Lange<sup>2</sup>, Augustyn Markiewicz<sup>6</sup>, Jan Peter Nap<sup>3</sup>, Jan van Oeveren<sup>7</sup>, Cyril Pommier<sup>8</sup>, Uwe Scholz<sup>2</sup>, Marco van Schriek<sup>7</sup>, Björn Usadel<sup>4,9</sup> and Stephan Weise<sup>2</sup>

[« Previous](#) | [Next Article »](#)  
[Table of Contents](#)

### This Article

*J. Exp. Bot.* (2015) 66 (18): 5417-5427.  
doi: 10.1093/jxb/erv271  
First published online: June 4, 2015  
This article appears in: [Special](#)

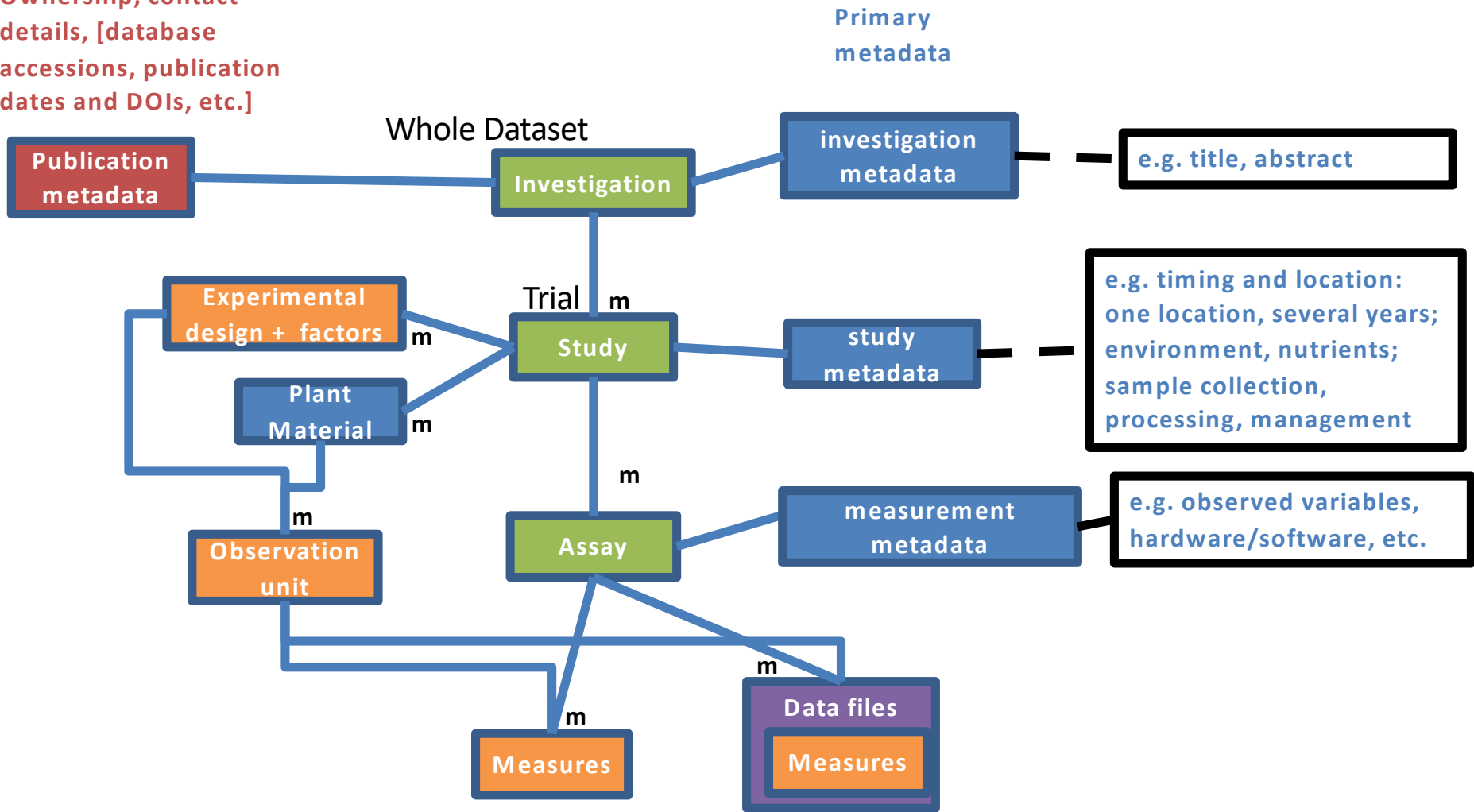
# MIAPPE Construction

- Contribution
  - ◆ European Infrastructures : Elixir (Bioinformatics), Emphasis/EPPN (Phenotyping)
  - ◆ National Institutes: France, Germany, Poland, UK, Portugal, Slovenia, Nederland, Belgium, Italy
- Steering committee
  - ◆ Elixir: European Bioinformatic infrastructure
  - ◆ Emphasis: European plant phenotyping infrastructure
  - ◆ Bioversity International - CGIAR
  - ◆ Elizabeth Arnaud, Paul Kersey, Pawel Krajewsky, Matthias Lange, Cyril Pommier, Bjorn Usadel
- Current Versions
  - ◆ Version 1 : Transplant + EPPN + Phenome
  - ◆ Version 1.1 : validation in progress
    - Documentation and model: Biologist friendly
    - Adaptations to new implementations



# MIAPPE v1.1 Overview

Ownership, contact details, [database accessions, publication dates and DOIs, etc.]



# MIAPPE v1.1 Overview

## Plant Material

- **Biosource : Plant Material identification**
- Multi Crop Passport Descriptor (MCPD) compliant
  - ◆ <https://www.bioversityinternational.org/e-library/publications/detail/faobioversity-multi-crop-passport-descriptors-v21-mcpd-v21/>
  - ◆ FAO & Bioversity International
  - ◆ Genebanks
- ID Fields are reused in MIAPPE
- Key : Identification
  - ◆ 0. Persistent unique identifier : PUID = URI/DOI
  - ◆ 1. Institute code
  - ◆ 2. Accession number
  - ◆ 5. Genus
  - ◆ 6. Species
  - ◆ 7. Subspecies : ‘subsp.’ (for subspecies); ‘convar.’ (for convariety); ‘var.’ (for variety); ‘f.’ (for form); ‘Group’ (for ‘cultivar group’)
  - ◆ 28. Remarks
- Suitable for non Genbank collections : breeder varietal list, laboratory collection, ...

## Variables

- Observation Variables
  - ◆ Phenotype & environment
  - ◆ Trait : What is measured
  - ◆ Method : How is it measured
  - ◆ Scale : How is it observed
  - ◆ Variable : trait + method + scale

traits, methods and scales

DOWNLOAD SHOW OBSOLETE TERMS English

- Wheat traits
  - Abiotic stress traits
  - Acid soil tolerance
  - AcidTol Estimation (method of)
  - Aluminum tolerance
  - Boron tolerance
  - Cold tolerance
  - Drought tolerance
  - Frost damage
  - Hail damage
  - Heat tolerance
  - Herbicide damage
  - Leaf rolling incidence

AcidTol Estimation

Identifier CO\_321:0000299

Method class Estimation

Method description Observe the growth of plants in low pH (acid) soil

Method name AcidTol Estimation

Method reference Reynolds et al., 2011

traits, methods and scales

search terms...

Woody Plant Ontology ONTOLOGY

Woody Plant Ontology ONTOLOGY

Ontology name Woody Plant Ontology

Authors Célia Michotey

Version v1

Licence CC BY 4.0

Links [CropOntology\\_TYDS format](#)  
[Adonis format](#)

- Physiological TRAIT CLASS
- Quality TRAIT CLASS
- Stress TRAIT CLASS
- Stress: Abiotic stress TRAIT CLASS
- Stress: Biotic stress TRAIT CLASS
- Bacterial canker girdling index TRAIT
- Bacterial canker lesion length TRAIT

NID: Nests presence VARIABLE

Identifier WOODY:0000067

Name NID

Synonyms Nests presence  
NID(Adonis)

Context of use Trial evaluation

Status Standard for INRA

Institution INRA

Scientist Célia Michotey

Date 13/03/2017

Crop Woody Plant

Nests presence TRAIT

Identifier WOODY:1000056

Name Nests presence

Description Presence or absence of nests

Main abbreviation NID

Status Standard for INRA

Class Stress: Biotic stress

Visual scoring METHOD

Identifier WOODY:2000003

Name Visual scoring

Description Visual assessment scoring scale

Class Estimation

Presence/ absence SCALE

# MIAPPE v1.1 overview:

## Metabolomic

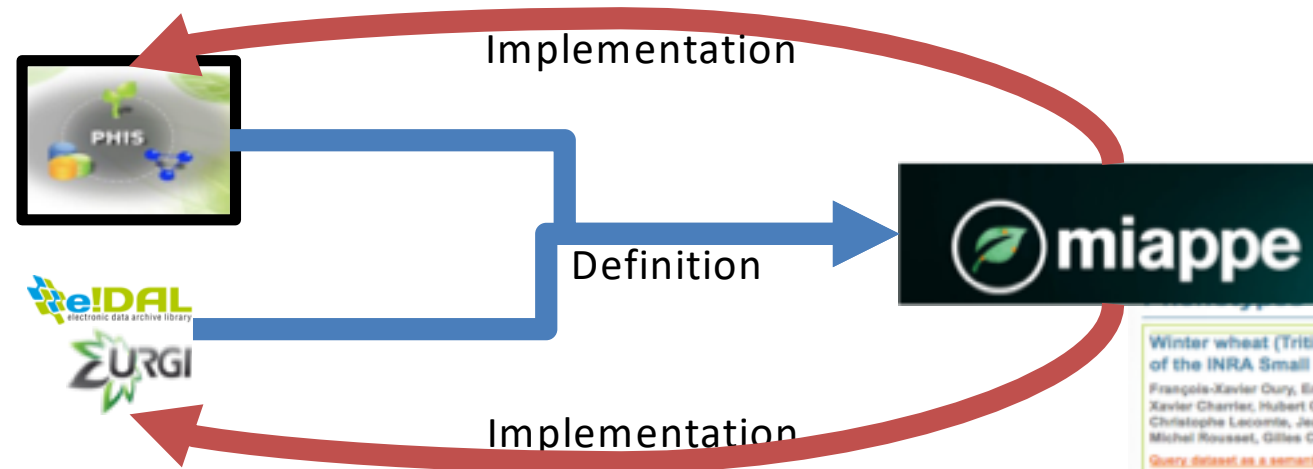
- Existing standards
  - ◆ Metabolights, Minimum information MSI
  - ◆ Full traceability
  - ◆ Sample treatment : temperature, tissue processing, molecule extraction protocol, measurement protocol....
- MIAPPE
  - ◆ Elaborated data
  - ◆ Elaborated, computed variables: Leaf Sugar content by microplot
  - ◆ Link to Metabo archive
    - Metabolight
    - Metabohub
    - ...

# MIAPPE DATA REPOSITORIES





# MIAPPE: Databases and Repositories



- Production databases
  - ◆ PHIS
- Cleaned data publication repositories
  - ◆ GnpIS
    - <http://dx.doi.org/10.15454/1.4489666216568333E12>
  - ◆ eDale
  - ◆ PlantPhenoDB
  - ◆ ...

Winter wheat (*Triticum aestivum* L.) phenotypic data from the multiannual, multilocal trial of the INRA Small Grain Cereals Network.

Frangola-Xavier Dury, Emmanuel Heumez, Bernard Rolland, Jérôme Auzanneau, Pierre Bérand, Maryse Brancourt, Xavier Charlier, Hubert Chiron, Camille Depatureaux, Laurent Falchetto, Olivier Gardet, Stéphane Gilles, Alex Girard, Christophe Lecomte, Jean-Yves Morlais, Pierre Pluchard, Didier Tropin, Maxime Trotlet, Patrice Walczak, Gérard Michel Rousselet, Gilles Charrel

Query dataset as a semantic graph.  
Or download the dataset as RDF archive.

Abstract  
Published 2015 by INRA

Back to Form

Search parameter(s):

Geolocation

DATA SETS: 4

Network Data Set :  
[INRA Wheat Network BPC accession \(A series\)](#)

Network Data Set :  
[INRA Small Grain Cereals Network](#)  
DOI:<http://dx.doi.org/10.15454/1.4489666216568333E12>

Network Data Set :  
[INRA Wheat Network not BPC accession \(B and C series\)](#)

Origin site Collecting site Evaluation site

Phenotyping campaign(s): 2000 x 2001 x 2002 x 2003 x 2004 x 2005 x 2006 x 2007 x 2009 x 2010 x 2011 x 2012 x 2013 x 2014 x 2015 x

remove all add all

Trial list Phenotypic data

Zurigo GnpIS B13 trials

Name	Campaign	Site	Geo 1 (lat)
B13: Champagne-côchelles_2004_Beta1	2004	Champagne-côchelles	
B13: Champagne-côchelles_2005_Beta1	2005	Champagne-côchelles	

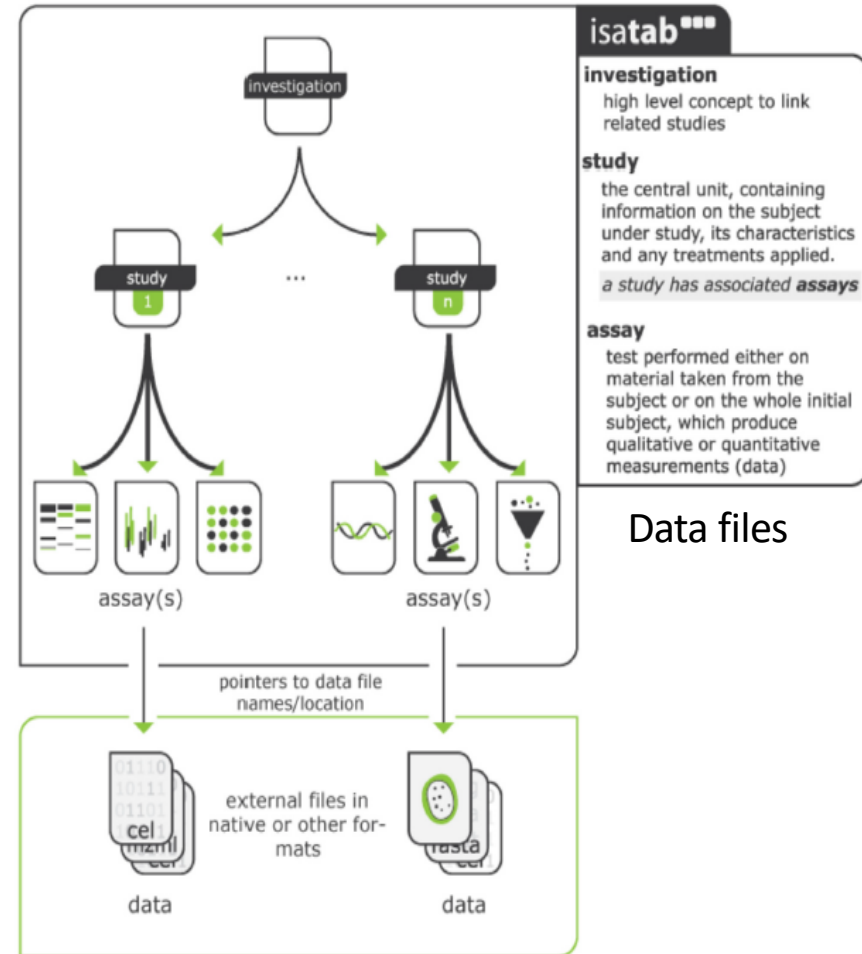
# MIAPPE FILE ARCHIVE



MIAPPE ISA Tab

# MIAPPE File Archive

- ISA Tab for Phenotyping
  - ◆ Investigation/Study/Assay
  - ◆ Zip Archive
    - MIAPPE Metadata
    - Raw data
      - CSV
      - Images or binary files
      - Reference to image archive (URI/URL)
    - Elaborated data
      - CSV
      - Provenance
- Training and improvements
  - ◆ File and metadata curation
  - ◆ BrAPI to IsaTab portable tool



## MIAPPE WEB SERVICES



- Breeding API
- International collaboration
  - ◆ Excellence in Breeding platform (CGIAR)
  - ◆ Coordinator : Peter Selby
  - ◆ Lead: Lukas Mueller, Jan Erik Backlund, Kelly Robbins
- Vision :
  - ◆ Standard Open API
  - ◆ Easy, secure, and efficient
  - ◆ Information Exchange
  - ◆ Between systems and applications
  - ◆ Breeding is the main target

Bill & Melinda Gates Foundation

CassavaBase

T3

IBP

JHI

Bioversity

CIRAD

INRA

IRRI

GOBII

Wageningen

CIP

DaRT

Cornell

iPlant



## MIAPPE STATUS AND PERSPECTIVES



# Adoption

- Plant community involved
  - ◆ Elixir (European bioinformatic infrastructure)
  - ◆ Emphasis (European Phenotyping infrastructure)
  - ◆ Bioversity international CGIAR
- Breeding API is Elixir official Phenotyping standard web service
- MIAPPE and BrAPI high collaboration
  - ◆ adoption and compliance
- Data repositories and management tools



- ◆ GnpIS <https://urgi.versailles.inra.fr/gnpis/>

- ◆ eDale <https://edal.iok-gatersleben.de/>

- ◆ PlantPhenoDB at IPGPAS <http://cropnet.pl/plantphenodb/>

- ◆ In progress: COPO, Elixir plant databases (iBet, WUR, VIB, ...), Brassica Information Portal, ...



# Perspectives

- Version 2 : Emphasis, Elixir, ...
  - ◆ Environment
  - ◆ Sensor traceability
  - ◆ Phenoharmonis Workshop Montpellier may 2018
- Elixir data lookup service, MIAPPE enabled.
  - ◆ Data discovery
  - ◆ Elixir
  - ◆ WheatIS & Emphasis ?
  - ◆ Open source software
- Dataset Validation
  - ◆ Elixir
  - ◆ File archive (ISA Tab) and BrAPI based





# Acknowledgment & Questions



◆ Paul Kersey



- IPG PAS
  - ◆ Hanna Cwiek-Kupczynska
  - ◆ Pawel Krajewski



- Bioversity international CGIAR
  - ◆ Elizabeth Arnaud
  - ◆ Marie Angélique Laporte



- ◆ Matthias Lange

- ◆ Cyril Pommier
- ◆ Anne Françoise Adam Blondon
- ◆ Guillaume Cornut
- ◆ Thomas Letellier
- ◆ Célia Michotey
- ◆ François Tardieu
- ◆ Pascal Neveu
- ◆ Manuel Ruiz
- ◆ Pierre Larmande



- ◆ Frederik Coppens



- iBet
  - ◆ Bruno Costa
  - ◆ Inês Chaves
  - ◆ Célia M. Miguel
- IGC
  - ◆ Daniel Faria



WAGENINGEN  
UNIVERSITY & RESEARCH

- ◆ Richard Finkers



- ◆ Bjorn Usadel