

➤ **Plant phenotyping data management from phenomics to integration for analysis and PGR characterizations: challenges and solutions from ELIXIR and EMPHASIS**

Solutions from ELIXIR and EMPHASIS European Infrastructure and beyond



EMPHASIS

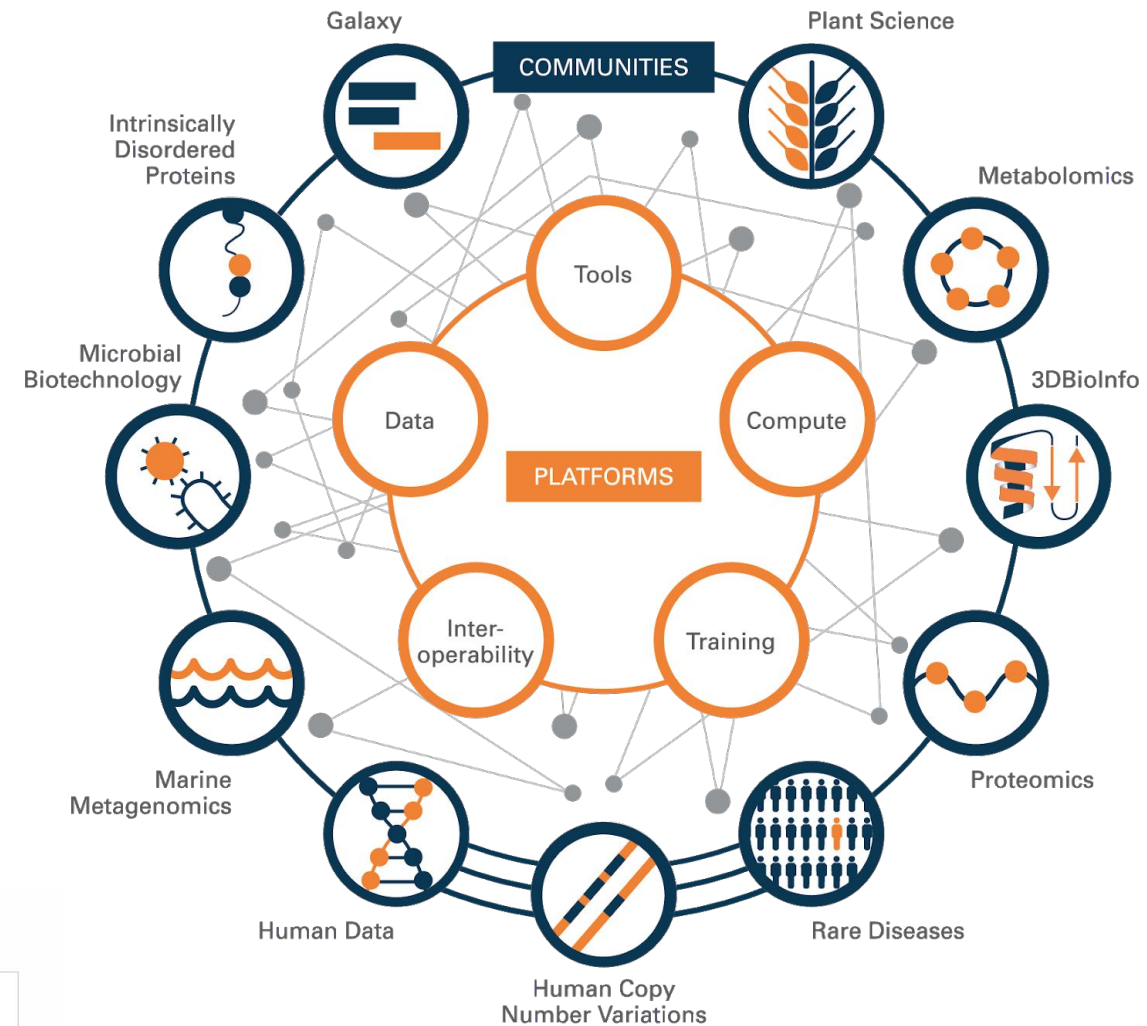
[www.elixir-europe.org](http://www.elixir-europe.org)

<https://emphasis.plant-phenotyping.eu>



# ELIXIR Plant community

- Sustainable tool federation
- FAIR data management
  - Adoption of standards in plant sciences
  - Develop community recommendations
  - Data portal
- Facilitate data integration and analysis
- Training.
- Joint projects (ELIXIR, EU, ...)



## Plant Sciences Community

### Leadership



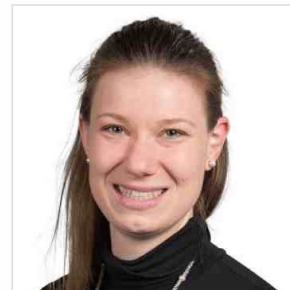
Sebastian Beier  
(ELIXIR Germany)



Kristina Gruden  
(ELIXIR Slovenia)



Cyril Pommier  
(ELIXIR France)



Katharina Heil  
(Communities Coordinator,  
ELIXIR Hub)



# Infrastructure Categories

PLANT PHENOTYPING REQUIRES INTEGRATION OF BOTH FACILITIES AND ACTIVITIES



CONTROLLED CONDITIONS

Investigation of diverse plant traits in response to well-defined environmental conditions



INTENSIVE FIELD

Detailed investigation of plants and canopies under well-monitored field conditions



LEAN FIELD

Field sites with basic equipment and environmental monitoring that can be linked to a network of field sites



MODELLING

Models integrated in phenotyping pipelines and predictive models using phenotypic data



DATA & COMPUTATIONAL SERVICES

Integrating compatible information systems to provide access to data



# Open science through FAIR data principles

Wilkinson et al., *The FAIR Guiding Principles for scientific data management and stewardship*. *Scientific Data* 3 (2016)

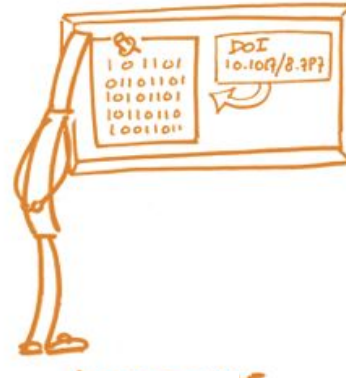
## FAIR DATA PRINCIPLES



**F**  
Findable



**Ids**  
**Index**  
**Metadata**  
**Description**



**A**  
Accessible



**Open Protocols**  
**Perennial Metadata**



**I**  
Interoperable



**Semantics**  
**Linked Data**  
**Vocabularies**



**R**  
Reusable



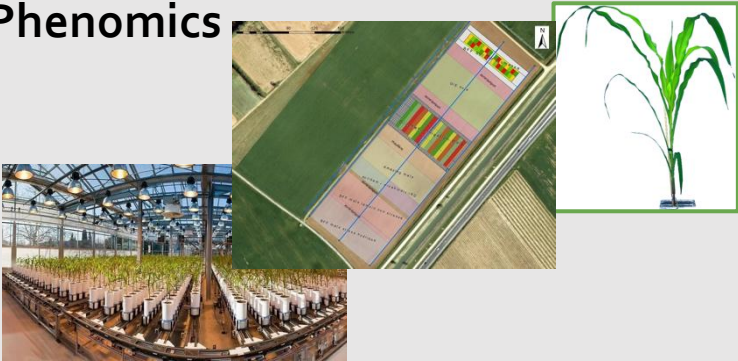
**License**  
**Well described**  
**Provenance (origin, process, methodology)**  
**Standards**

Sustainable data access over decades

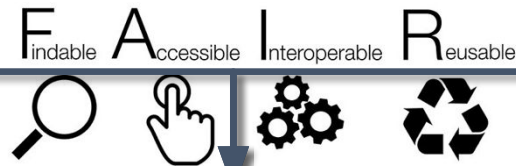
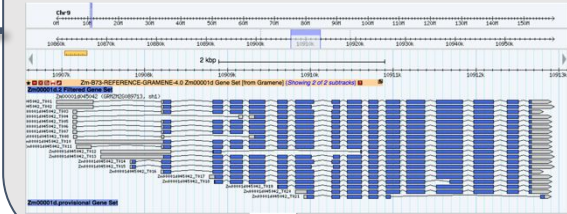
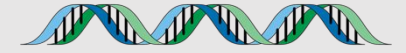
# PLANT use case

- Environment / Phenome / Genomic / \*omic / Genetic

## Phenomics



## Genetics Genomics Omics



**Plant Breeding**  
Genetic variations by Traits

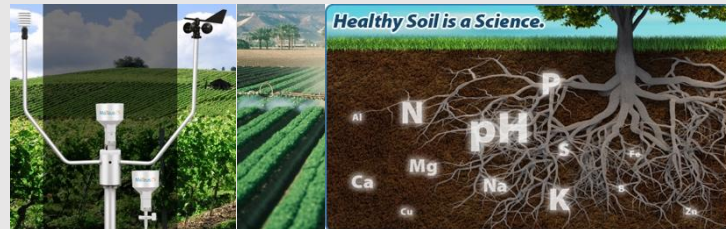
**Climate Change Studie**  
Genotype by Environment

Dispersed  
Heterogenous  
Getting Standardized

Mostly centralized  
Homogenous data  
Heterogenous metadata



## Environment



Dispersed  
Heterogenous

# • What is FAIR for plant data ?

## • Phenotyping

- Raw data
  - Images in different modes
    - RGB / spectral / hyperspectral / thermal
  - Individual plant time series
  - Expensive to generate
  - Not reproducible
- Computed / derived data
  - Data matrices (XLSX)

## • Genetic variation

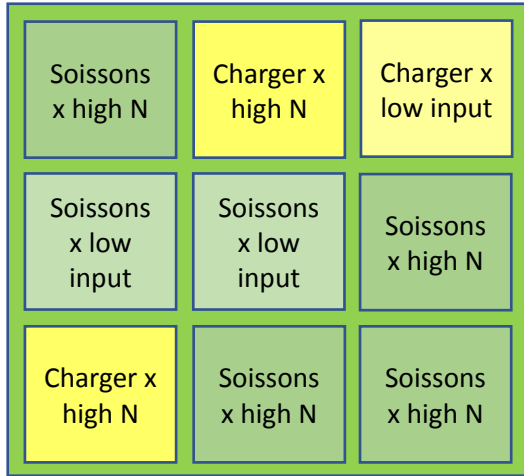
- Raw data
  - Sequence files
  - “cheap” to generate
  - Big Data
- Derived
  - VCF
  - Aligned to a given reference genome

## • Envirotyping

- Raw data
  - numeric
  - highly dynamic time series
  - spatial heterogeneity

# Phenotyping data life cycle

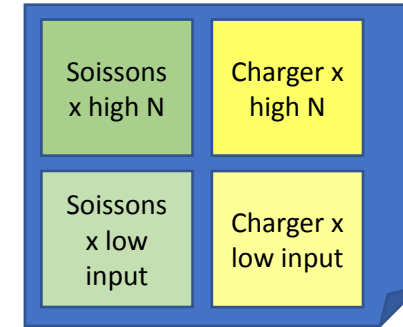
« Raw » data, pheno/env measures, variables



Derivation, Reduction

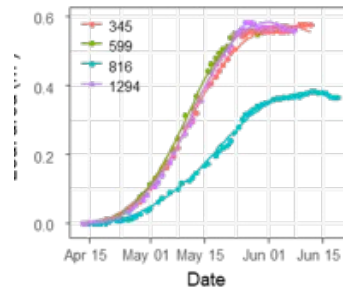


« computed » data, reduced, indicators



Genotype	Treatment	N input	Date	Rep	Fusariose
Soissons	low input	15,32253129	15/11/2011	1	5
Soissons	low input	15,31430556	16/11/2011	2	7

Genotype	Treatment	Fusariose
Soissons	low input	6



661300270	Ardon	2004	45.645632645603683	12/01/2004	284.3
661300270	Ardon	2005			
661300444	Ardon	2004	38.96112577281653	12/01/2004	228.8
661300444	Ardon	2005			
661300312	Cavallermaggiore	2004	52.4	01/01/2004	249.9
661300312	Cavallermaggiore	2005			
661300371	Cavallermaggiore	2004	45.74	01/01/2004	230.2
661300371	Cavallermaggiore	2005			
661300487	Cavallermaggiore	2004	72.52	01/01/2004	309.8
661300487	Cavallermaggiore	2005			
661300585	Cavallermaggiore	2004	71.73999999999999995	01/01/2004	305.7
661300585	Cavallermaggiore	2005			
661300468	Headley	2004	45.27	01/01/2004	
661300468	Headley	2005			
661300469	Headley	2004	70.93000000000000007	01/01/2004	
661300469	Headley	2005			
661300533	Headley	2004	57.67	01/01/2004	258.8

# Plant Phenotyping Life cycle

## Raw data long term conservation

### Data acquisition

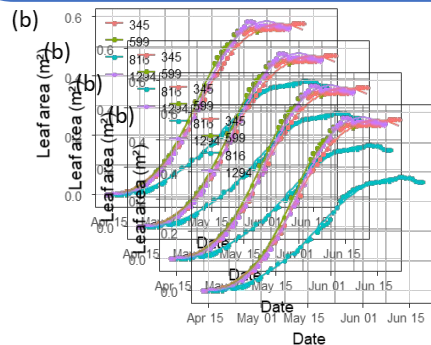
- **VARIABLES**
- Plant/microplot level
- Traceability
- Raw measures
- Data Cleaning
- Platform IS (Emphasis IS, PHIS, ...)
- Analysis Reproducibility
- Provenance

### Data computation

- **INDICATORS**
- Statistical integration
- Genotype level (mostly)
- New computation for each scientific question
- One raw dataset  many computed datasets

### Data publication

- One Data Publication by datasets.
- **Platform IS**
  - Phenomic, plant level
- **FAIR Data Repositories**
  - Reduced



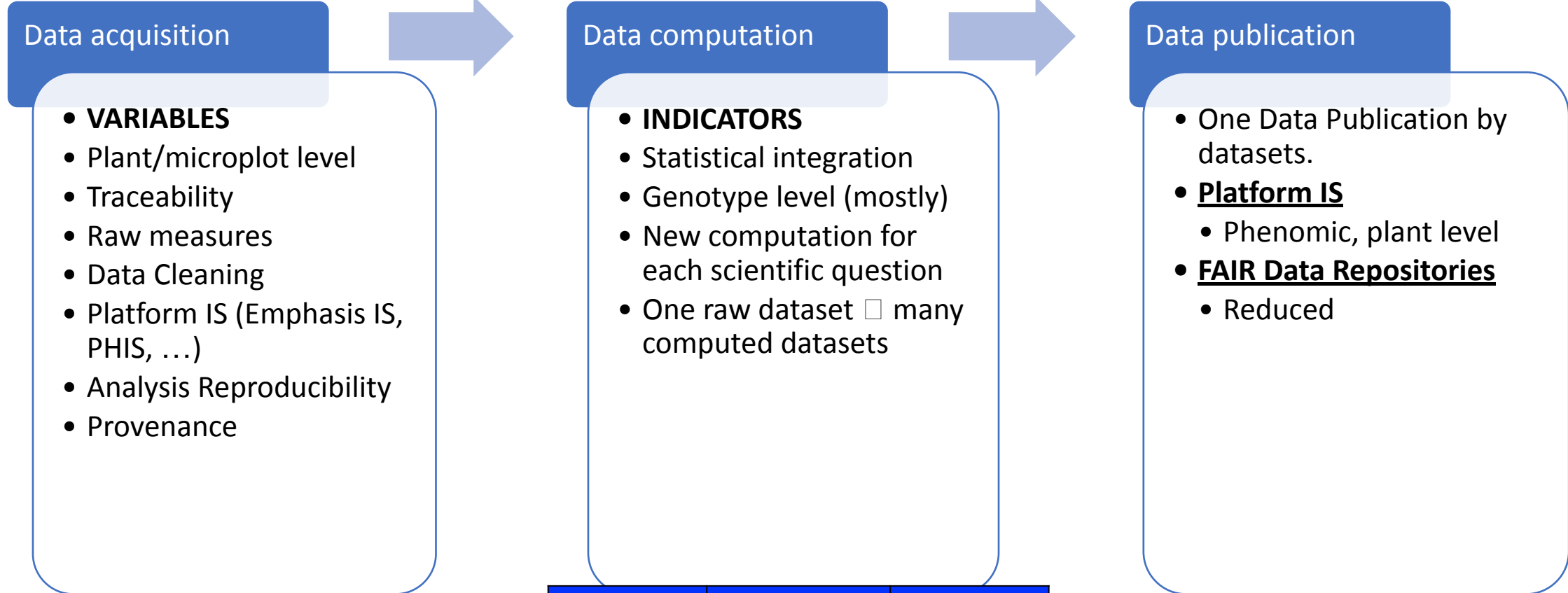
Genotype	traitement	Fusariose
Soisson	low input	5
Soisson	high N	7
Charger	low input	1
Charger	high N	2

Variety charger is resistant to fusariose under intensiv cultural practice



# Plant Phenotyping Life cycle

## Raw data long term conservation



### Data acquisition

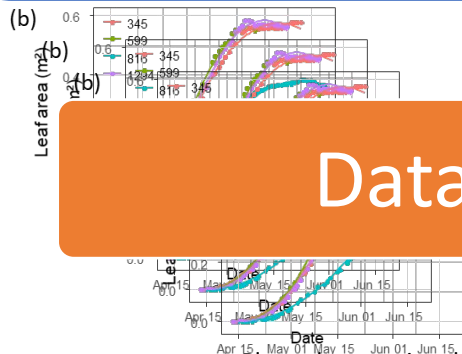
- **VARIABLES**
- Plant/microplot level
- Traceability
- Raw measures
- Data Cleaning
- Platform IS (Emphasis IS, PHIS, ...)
- Analysis Reproducibility
- Provenance

### Data computation

- **INDICATORS**
- Statistical integration
- Genotype level (mostly)
- New computation for each scientific question
- One raw dataset  many computed datasets

### Data publication

- One Data Publication by datasets.
- **Platform IS**
  - Phenomic, plant level
- **FAIR Data Repositories**
  - Reduced



**Data**

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

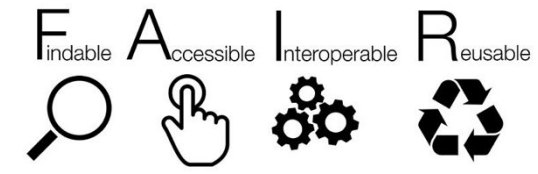
**Knowledge**

Variety charger  
intensive cultural practice

Plant phenotyping data management from phenomics to integration for analysis and PGR characterizations: challenges and solutions from ELIXIR and EMPHASIS

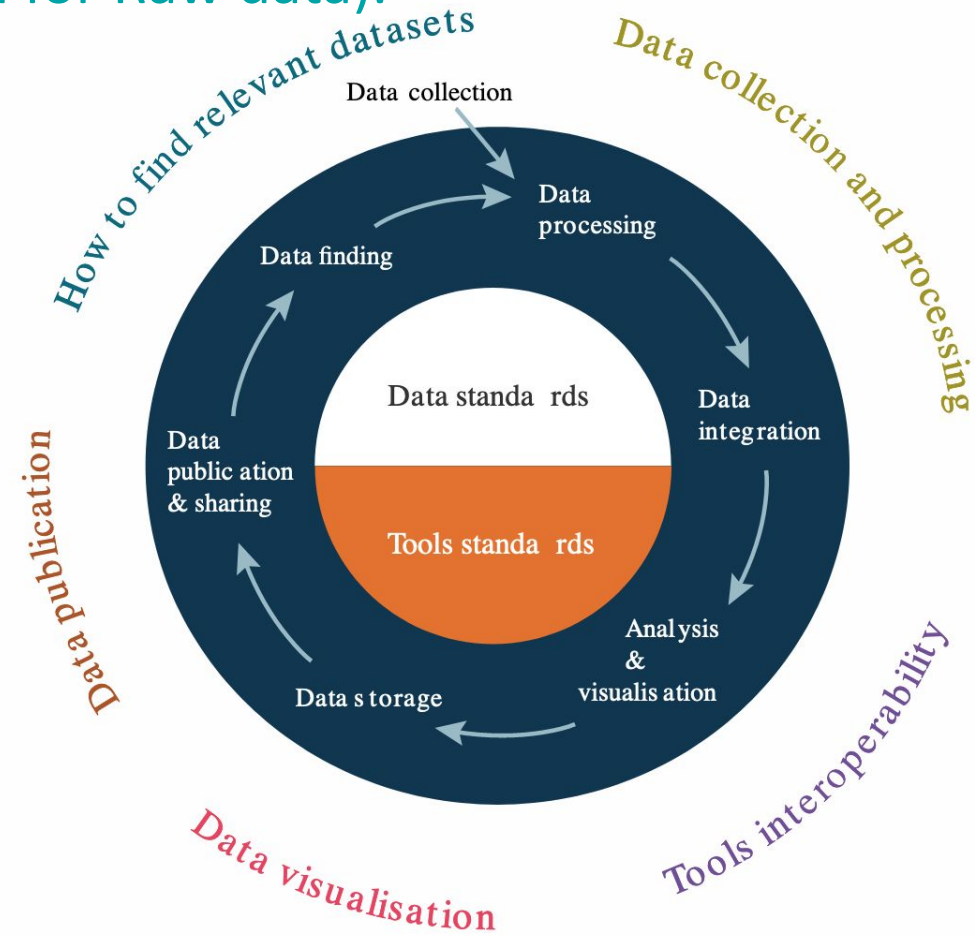
## • Plant Genetic variation

- Variability of the genotypes (AKA varieties, accessions, germplasm)
- Sequencing (GBS), Chips, ...
- Raw data : reads
- Aligned data : VCF
- Paradigm: Raw data is too big, easy to generate  keep only Variation
- But: realign to a new genome version, or to another reference variety ?
- Raw data can be interesting to keep too



# FAIR For plant science

- Phenotyping: Raw and derived data
- Genotyping: Computed data (plus option for Raw data).
  - Applies to other OMICS
- Solutions on the data lifecycle
  - Data standardisation
  - Data repositories for publication
  - Data findability / discovery

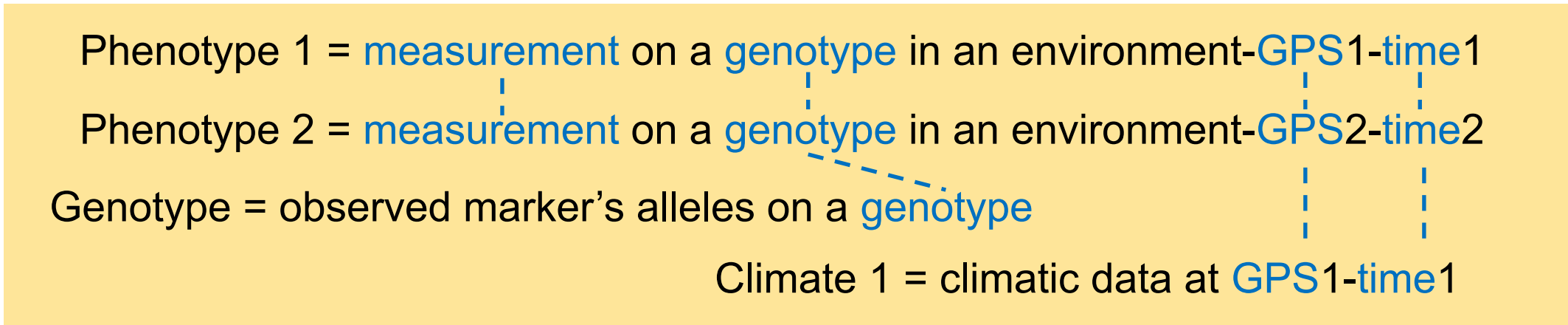


- **PLANT DATA STANDARDS : WHY**



# Why should we standardize data?

- Allow anyone (including yourself) to reuse it: **metadata about the experiment (who did it, for what purpose, where and how)**
- Enable data integration with other types of data: **Linked data between datasets using identification of pivot objects**



- To enable knowledge discovery: **metadata about the experiment, controlled vocabularies, ontologies**

- **PLANT DATA STANDARDS : WHO**

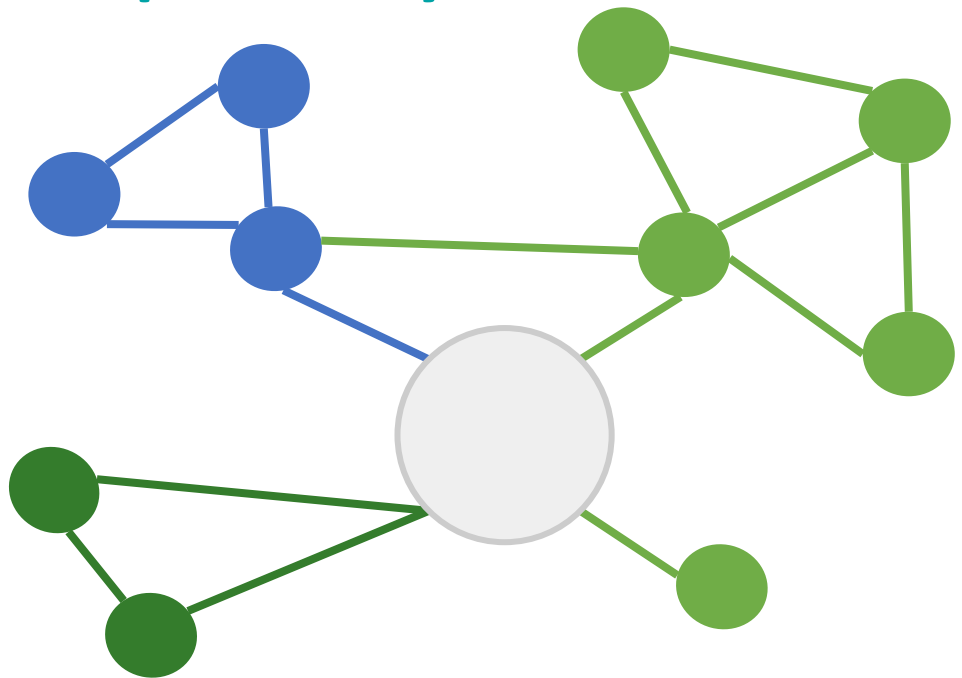


- Interoperability in International network

**National Networks**



**Global Networks**



**European Networks**



**International data standards**

- **Sharing standards: standards registries**

The screenshot displays the FAIRsharing.org interface. At the top left is the logo for FAIRsharing.org, which includes the text "standards, databases, policies". To the right of the logo is a search bar with the placeholder text "Search all of FAIRsharing". Further right are navigation buttons for "Standards", "Databases", "Policies", "Collections", "Add/Claim Content", "Stats", and "Log".

The main content area contains three informational boxes:

- How to cite this record** FAIRsharing.org: GnpIS; 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](#).

Below these boxes are two columns of links:

- In Collections**
  - [Wheat Data Interoperability Guidelines](#)
  - [ELIXIR node contributed resources](#)
- 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](#)

At the bottom left, there are sections for "Support" and "General".



# Community driven recommendations and registries

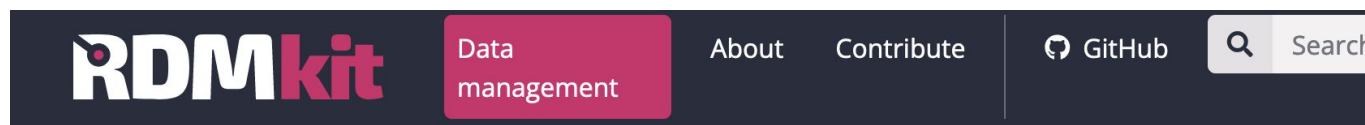
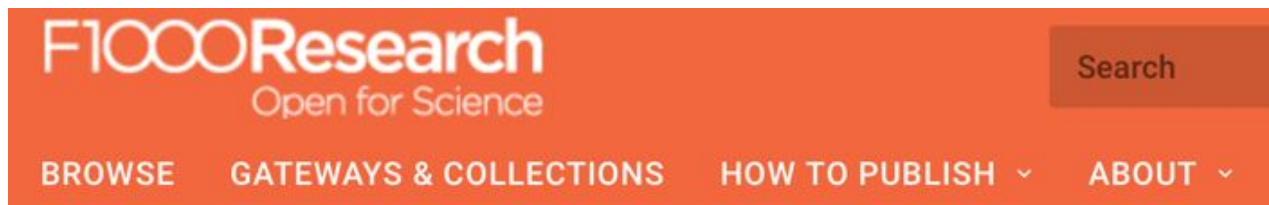
- WheatIS:

<http://wheatis.org/DataStandards.php>

- RDM Toolkit

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

- Community story



## ARTICLE Developing data interoperability using standards: a community use case [version 2; referees: 2]

Developed and managed by people who work every day with life science data, RDMkit has guidelines, information, and pointers to help you with problems throughout the life cycle. RDMkit supports FAIR data — Findable, Accessible, Interoperable, Reusable — by-design, from the first steps of data management planning to the final storage of data in public archives.

### Are you working with data in the Life Sciences? Do you feel overwhelmed when you think about Research Data Management?

The ELIXIR Research Data Management Kit (RDMkit) is an online guide containing management practices applicable to research projects from the beginning to the end. Developed and managed by people who work every day with life science data, RDMkit has guidelines, information, and pointers to help you with problems throughout the life cycle. RDMkit supports FAIR data — Findable, Accessible, Interoperable, Reusable — by-design, from the first steps of data management planning to the final storage of data in public archives.

The RDMkit organises information into the six sections displayed below, which are interconnected but can be browsed independently.

Yeumo<sup>1</sup>, Michael Alaux <sup>2</sup>, Elizabeth Arnaud<sup>3</sup>, Sophie Aubin<sup>1</sup>, Ute Baumann<sup>4</sup>, the<sup>5</sup>, Laurel Cooper <sup>6</sup>, Hanna Ówiek-Kupczyńska<sup>7</sup>, Robert P. Davey <sup>8</sup>, dan Fulss<sup>9</sup>, Clement Jonquet <sup>10,11</sup>, Marie-Angélique Laporte<sup>3</sup>, Pierre Larmande <sup>12,13</sup>, nier <sup>2</sup>, Vassilis Protonotarios <sup>14</sup>, Carmen Reverte <sup>15</sup>, Rosemary Shrestha<sup>9</sup>, rats<sup>16</sup>, Aravind Venkatesan <sup>12</sup>, Alex Whan<sup>17</sup>,  Hadi Quesneville <sup>2</sup>

details  
This article is included in the [Global Open Data for Agriculture and Nutrition gateway](#).

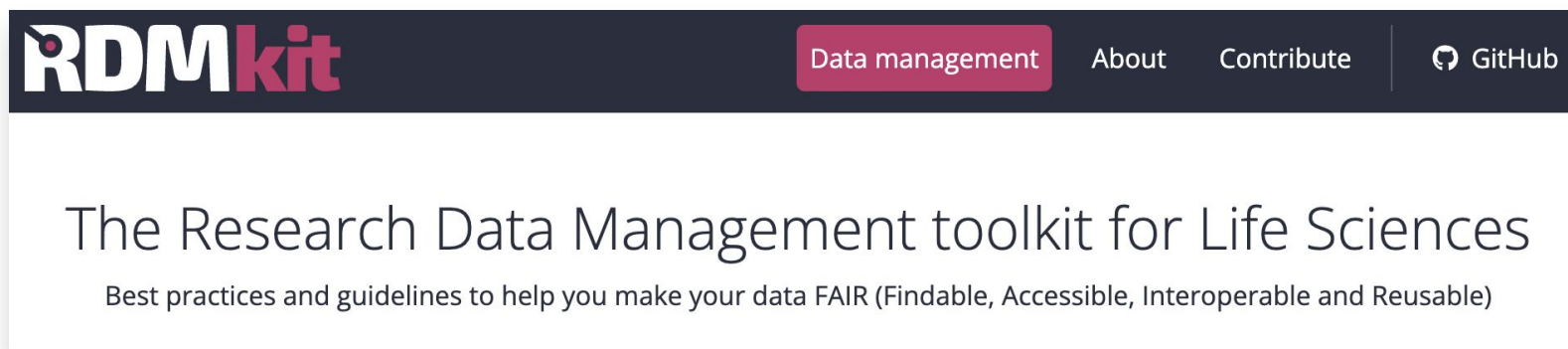
- Data management
- Data life cycle
- Your role
- Your domain
- Your tasks
- Tool assembly
- National resources
- All tools and resources
- All training resources

### Data life cycle



# Community guidelines portal : RDMkit - Best practices and guidelines for FAIR data management

- A “wikipedia-like” knowledge base website, free and open
- Describes how to manage research outputs according to FAIR principles
- Portal to other online resources used by RDM professionals and researchers

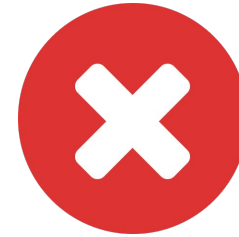


URL: <https://rdmkit.elixir-europe.org/>

Recommended in the **Horizon Europe Program Guide** as the “resource for Data Management guidelines and good practices for the Life Sciences”

# RDMkit is NOT

- A peer reviewed journal
- A repository to store files
- A registry for listing resources and tools
- A manual or user guide for tools
- A website for projects to share deliverables



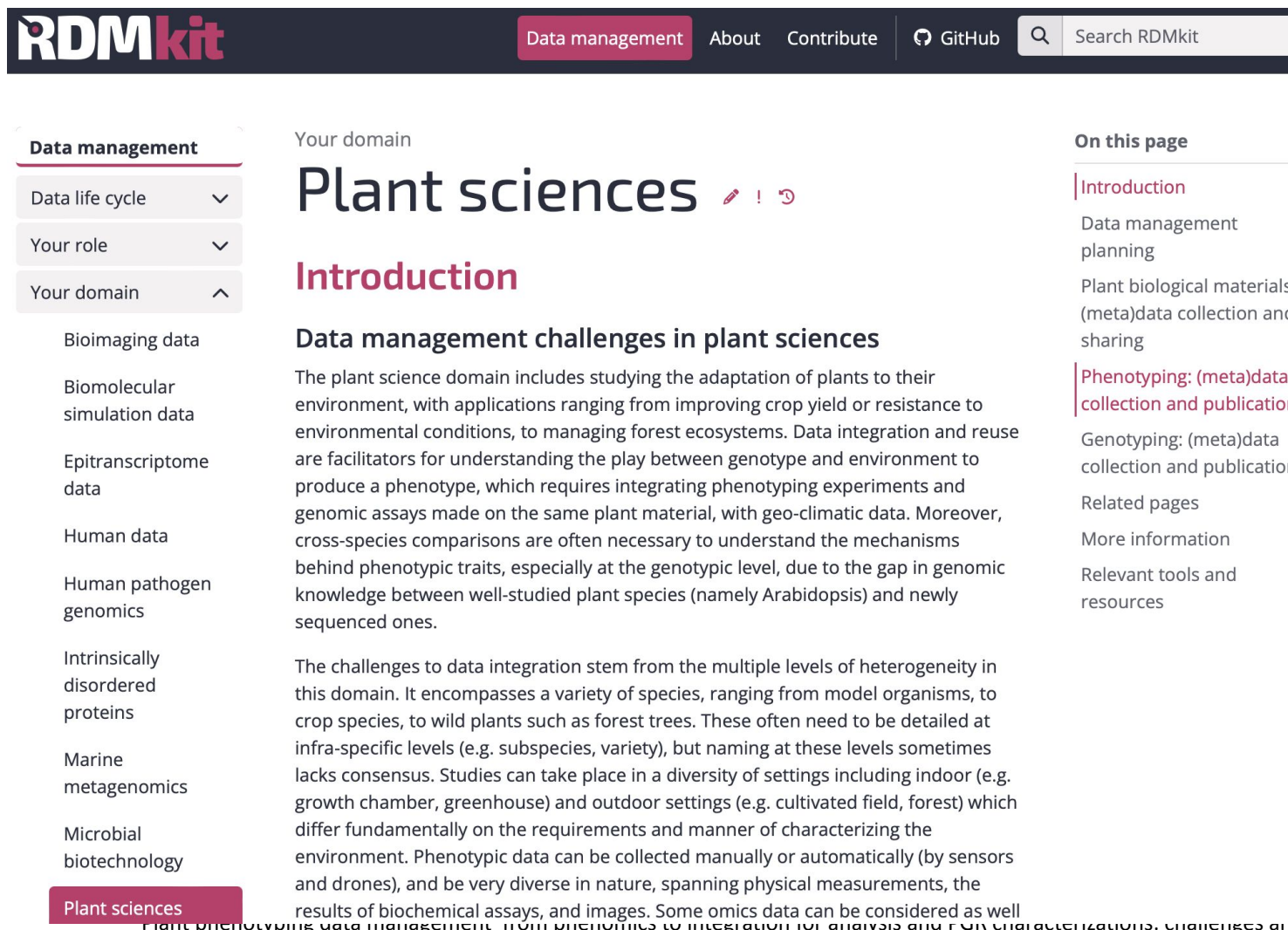
# RDMkit in numbers



- Contributors are experts in RDM and/or in scientific domain in ELIXIR and beyond
- ELIXIR Node service in 3 Nodes/Countries
- Managed by an editorial board composed of 16 ELIXIR members from several Nodes and the lead for “FAIR Data & Resources” at the Office of Data Science Strategy at NIH
- The content is curated by members of the ELIXIR RDM Community

# Plant Pages including PGR

[https://rdmkit.elixir-europe.org/plant\\_sciences](https://rdmkit.elixir-europe.org/plant_sciences)



**RDMkit** Data management About Contribute GitHub Search RDMkit

**Data management**

- Data life cycle
- Your role
- Your domain

Bioimaging data

Biomolecular simulation data

Epitranscriptome data

Human data

Human pathogen genomics

Intrinsically disordered proteins

Marine metagenomics

Microbial biotechnology

**Plant sciences**

Your domain

## Plant sciences

### 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

**On this page**

- Introduction
- Data management planning
- Plant biological materials: (meta)data collection and sharing
- Phenotyping: (meta)data collection and publication
- Genotyping: (meta)data collection and publication
- Related pages
- More information
- Relevant tools and resources

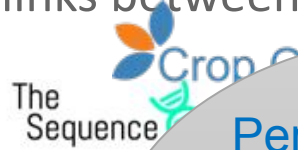
- **PLANT DATA STANDARDS : WHAT**



# • Data standards for FAIR

## Semantic

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



## Persistent Unique Identifiers

URI, gene ID, accessions ID, Trait ID, DOI,...

## Structure

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



## Technical

- Data integration and sharing
- Interoperability : tools and systems
  - GA4GH
  - Breeding API [www.brapi.org](http://www.brapi.org)
  - Computer scientist driven



# • Semantic Standard: Ontologies

## • Annotating one object

- Protein, gene
- Plant, plant anatomy, ...

## • Atomic concept

- protein function
- cellular localization
- ...

UniProtKB - P56761 (PSBD\_ARATH)

Display [Help video](#) [BLAST](#) [Align](#) [Format](#) [Add to basket](#) [History](#)

Entry **Protein** **Photosystem II D2 protein**

Publications **Gene** **psbD**

Feature viewer **Organism** *Arabidopsis thaliana (Mouse-ear cress)*

Feature table **Status** Reviewed - Annotation score: ●●●●●● - E:

None **Function!**

Gene Product	Symbol	Qualifier	GO Term	Evidence	Reference	With / From	Taxon
UniProtKB:P56761	psbD	is_active_in	GO:0009535	ECO:0000318 IBA	PMID:21873635	PANTHER:PTN002108145 more...	3702 Arabidopsis thaliana
UniProtKB:P56761	psbD	part_of	GO:0009523	ECO:0000318 IBA	PMID:21873635	PANTHER:PTN002108145 more...	3702 Arabidopsis thaliana
UniProtKB:P56761	psbD	enables	GO:0005515	ECO:0000353 IPI	PMID:25846821	UniProtKB:Q9FL44	3702 Arabidopsis thaliana
UniProtKB:P56761	psbD	involved_in	GO:0019684	ECO:0000256 IEA	GO_REF:0000002	InterPro:IPR000484 more...	3702 Arabidopsis thaliana



- Semantic Standard: Ontologies for Phenotype

- 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

# Ontologies, variables, descriptors

- Phenotyping variables
  - Crop Ontology
    - Collection of species specific ontologies
    - <https://www.croponontology.org/>
    - Plus <https://urgi.versailles.inra.fr/ontologyportal>
    - Methods can be specific of a community (CGIAR, INRAE)
  - Consortium dedicated ontology
  - Or contribution to the ontologies
  - Little to no PGR Descriptors
- PGR Descriptors
  - Might be added in Crop ontologies (e.g. Grape)
  - IPGRI descriptors
  - Rarely in ontologies
- Ontology term search engine
  - <https://www.ebi.ac.uk/ols4>

The screenshot shows the EBI OLS4 ontology search interface. On the left, a search bar contains 'irms...'. Below it, a list of traits is displayed, including 'Biotic stress' (TRAIT CLASS), 'Morphological' (TRAIT CLASS), and various specific traits like 'Berry hilum', 'Berry pedicel', 'Berry: bloom', 'Berry: color of skin', 'MORPHO\_OIV\_225: Berry: color of skin', 'Berry: ease of detachment from pedicel', 'Berry: firmness of flesh', 'Berry: formation of seeds', 'Berry: intensity of the anthocyanin coloration of fl', 'Berry: juiciness of flesh', 'Berry: length', 'Berry: length of seeds', 'Berry: particularity of flavor', and 'Berry: shape'. The 'MORPHO\_OIV\_225: Berry: color of skin' trait is selected and highlighted in blue. On the right, a detailed view of this trait is shown, including its ontology name, identifier, name, synonyms, institution, scientist, date, crop, and other metadata.

Ontology name	Vitis-inra-ontology
Identifier	CO_356:1000017
Name	MORPHO_OIV_225
Synonyms	Berry: color of skin
Institution	INRAE
Scientist	Eric Duchene
Date	00/00/00
Crop	VITIS

Berry: color of skin	TRAIT
Identifier	CO_356:2000007
Name	Berry: color of skin
Entity	Berry skin
Attribute	Color
Class	Morphological

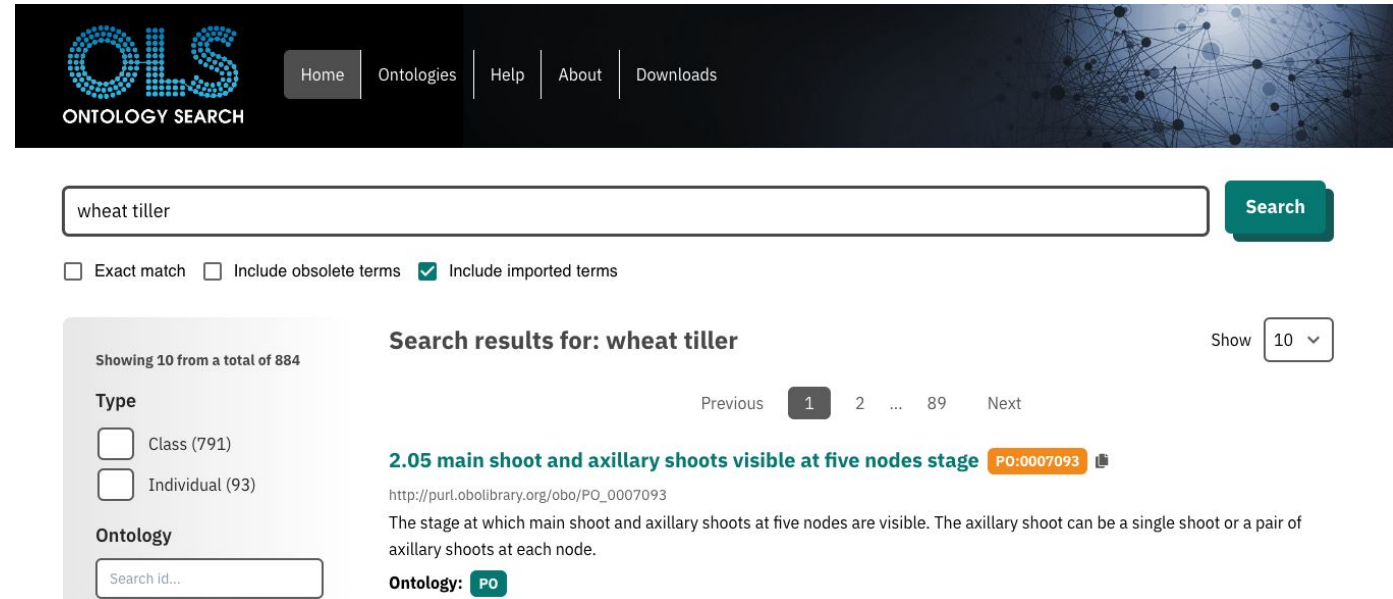
OIV:225	METHOD
Identifier	CO_356:3000077
Name	OIV:225
Description	Berry: color of skin
Class	Estimation

S1_6	SCALE
Identifier	CO_356:4000031
Name	S1_6
Data type	Nominal
Decimal places	0
Min	1
Max	6

# Ontologies, variables, descriptors

- Phenotyping variables
  - Crop Ontology
    - Collection of species specific ontologies
    - <https://www.croponontology.org/>
    - Plus <https://urgi.versailles.inra.fr/ontologyportal>
    - Methods can be specific of a community (CGIAR, INRAE)
  - Consortium dedicated ontology
  - Or contribution to the ontologies
  - Little to no PGR Descriptors
- PGR Descriptors
  - Added in Crop ontologies (e.g. Grape)
- Ontology term search engine
  - <https://www.ebi.ac.uk/ols4>



The screenshot shows the OLS (Ontology Lookup Service) search interface. At the top, there is a navigation bar with links for Home, Ontologies, Help, About, and Downloads. The search bar contains the text "wheat tiller". Below the search bar, there are checkboxes for "Exact match", "Include obsolete terms", and "Include imported terms". The search results are displayed in a list format, showing "Showing 10 from a total of 884". The first result is "2.05 main shoot and axillary shoots visible at five nodes stage" with the ID "PO:0007093". The description of the result is "The stage at which main shoot and axillary shoots at five nodes are visible. The axillary shoot can be a single shoot or a pair of axillary shoots at each node." The ontology is identified as "PO".

- **Phenotype Structure Standard**



## Minimal Information About Plant Phenotyping Experiment : version 1.1 (Jan 2019)

[www.miappe.org](http://www.miappe.org)

- **Many stakeholders**

- ◆ Elixir, Emphasis, Bioversity, North American PPN

- **Open Community:**

- ◆ Request for comments
- ◆ Github Feature requests
- ◆ Mailing lists
- ◆ Meetings & Workgroups

- **Crops and woody plants**

Papoutsoglou *et al.* (2020) Enabling reusability and interoperability of plant phenomic datasets with MIAPPE 1.1. *New Phytol*, 227:260-273; <https://doi.org/10.1111/nph.16544>

MIAPPE					
line #	MIAPPE Check list	Definition	Example	Format	Cardinality
DM-1	<b>Investigation</b>	Investigations are research programmes with defined aims. They can exist at various scales (for example, they could encompass a grant-funded programme of work, the various components comprising a peer-reviewed publication, or a single experiment).			1 per MIAPPE submission
DM-2	<b>Investigation unique ID</b>	Identifier comprising the unique name of the institution/database hosting the submission of the investigation data, and the accession number of the investigation in that institution.	EBI12345678	Unique Identifier	0-1
	<b>Investigation title</b>	Human-readable string summarising the investigation.	Adaptation of Maize to Temperate Climates, Mid-Density Genome-Wide Association Genetics and Diversity Patterns Reveal Key Genomic Regions, with	Free text (short)	1
Environment					
ENV-1	Non-exhaustive list of Environment Parameters.				
ENV-2	Environment parameters	Definition	Example	Format	
ENV-3	Growth facility				
ENV-4	<b>Air temperature</b>	List of hourly air temperature throughout the experiment.	22 °C	Numeric	
ENV-5	<b>Organ temperature</b>	List of hourly organ temperatures throughout the experiment.	18 °C	Numeric	
Experimental Factors					
TR-1	Non-exhaustive list of Experimental Factors that can be applied.				
TR-2	Factor type	Definition	Example factor values	Format	
TR-3	<b>Seasonal environment</b>	A plant treatment (EO:0001001) involving an exposure to a given conditions of regional seasons.	Spring season; dry season	Plant Environment Ontology:'EO_0007038'	
TR-4	<b>Air treatment regime</b>	The treatment involving an exposure to wind/air with varying degree of temperature, which may depend on the study type or the regional environment.	28/25°C ( Day/Night )	Plant Environment Ontology:'EO_0007161'	
TR-5	<b>Soil temperature regime</b>	A physical plant treatment (EO:0007316) involving an exposure to varying degree of temperature, which may depend on regional environment.	27/25°C ( Day/Night )	Plant Environment Ontology:'EO_0007161'	

- **Phenotype Structure Standard**



Minimum Information for Biological and Biomedical Investigations

A collection of the historical MIBBI foundry reporting guidelines. The minimum information standard is a set of guidelines for reporting data derived by relevant methods in biosciences. If followed, it ensures that the data can be easily verified, analysed and clearly

- **Biologist Friendly**

- Clear definitions and examples
- Excel templates
- Trainings

- **Minimal and sufficient list of metadata:**

- The objective of the experiment
- Who contributed to the experiment
- What were the experimental procedures
- What was the biological material experimented
- ...

# • Phenotype Technical Standard, MIAPPE Implementations

## • Ontology, OWL Implementation

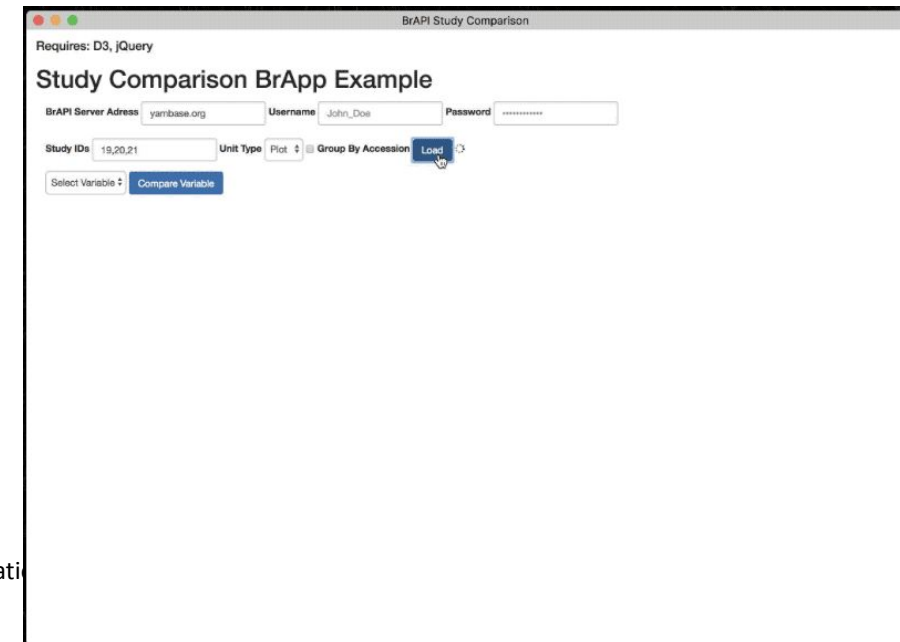
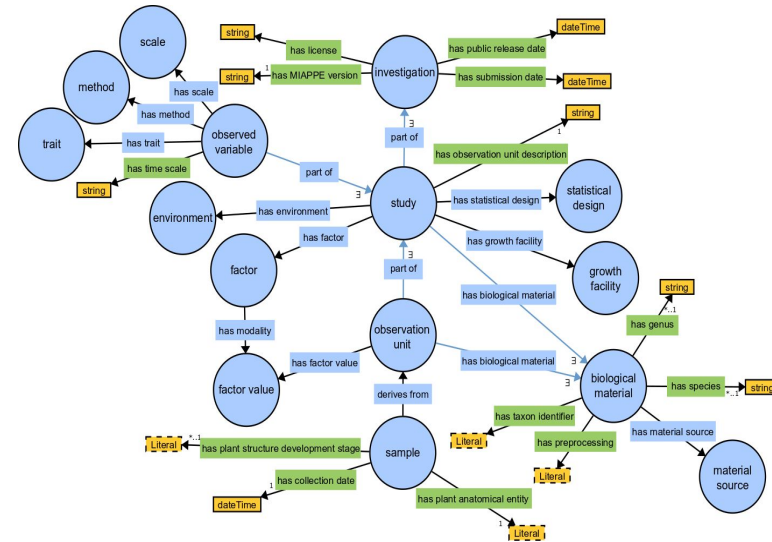
- <https://github.com/MIAPPE/MIAPPE-ontology>
- <http://agroportal.lirmm.fr/ontologies/PPEO>
- Data model representation
- Formal concepts and constraints

## • File Archive

- ISA Tab: data + metadata
- RO Crate studies

## • Web Services

- Breeding API
- International collaboration
- Standard Open Web Service API
- Information Exchange, Main target: Breeding
- Excellence in Breeding platform (CGIAR, Peter Selby)

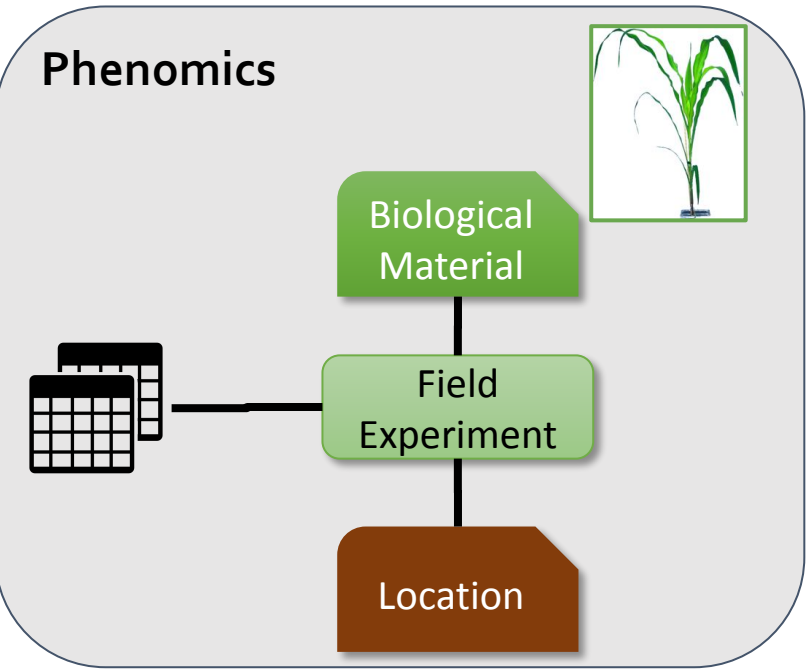


# Data management tools

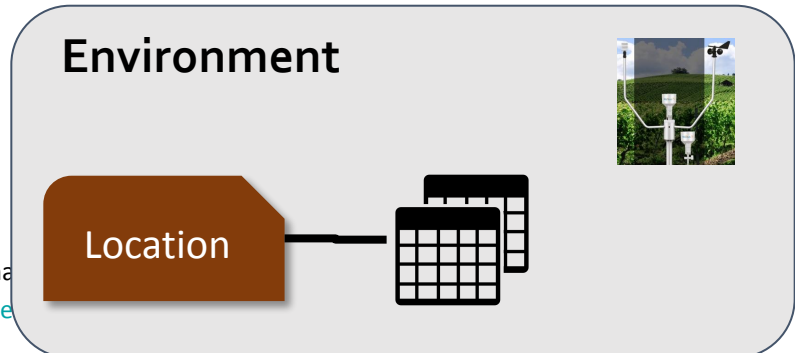
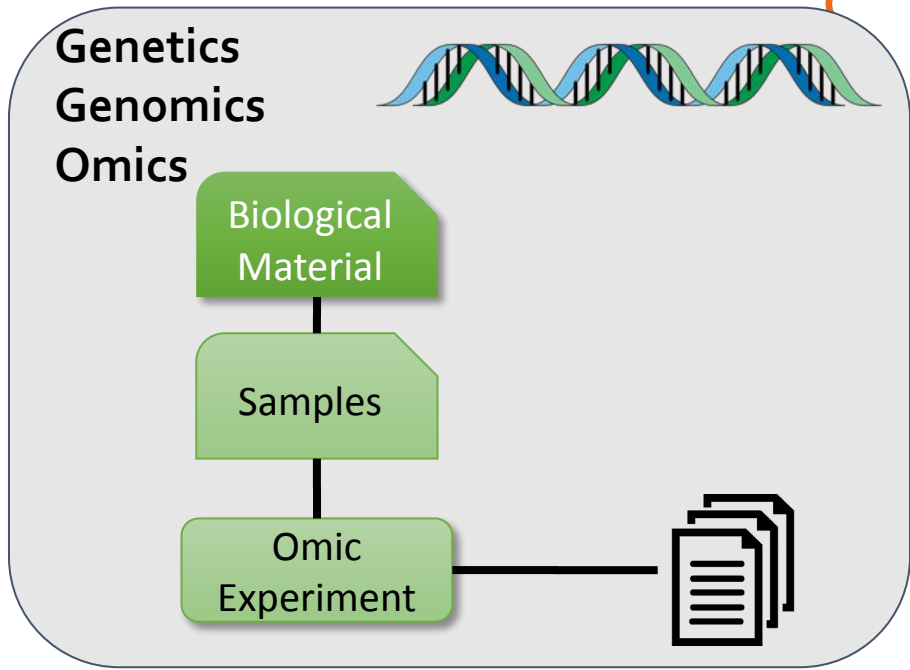
- Simple phenotyping network
  - Fairdom Seek (AGENT H2020)
  - Simple shared space for data, files SOP, etc..
  - Avoid “sent by email”
  - **Usable beyond Phenotyping**
- Fully instrumented Phenotype
  - Emphasis solutions (PIPPA, PHIS, IPK, ...)
    - <https://emphasis.plant-phenotyping.eu/services/emphasis-pilots/data-services>

The screenshot shows the 'WP3-Phenotypic-Historical' project page. At the top, there are tabs for 'Overview' and 'Related items'. Below the tabs, a message states: 'Please use the following template: [Eurisco Phenotyping Historical data template](#)'. It also refers to a 'Helpdesk FAQ' for data management in FAIRDOM. A list of three required traits is provided: 'plant height, written "Plant height"', 'flowering time, either as "Days to heading" (Type: Measurement), or "Date of heading" (Type: Date)', and 'thousand-kernel weight, written "Thousand Kernel Weight"'. Below this, the 'SEEK ID' is given as <https://urgi.versailles.inrae.fr/fairdom/investigations/2>, and the 'Projects' are listed as 'H2020-AGENT'. The 'Investigation position' is noted as '4'. A 'Selected' section highlights the current investigation, providing a 'Description' that repeats the template instruction and the same 'SEEK ID'. At the bottom, a hierarchical tree view shows the project structure: 'WP3-Phenotypic-Historical' (selected) contains 'CREA-CI-Phenotypic Historical Study wheat', which includes 'CREA\_Phenotypic historical data wheat\_01-2023' (containing 'CREA Phenotypic hystorical assay\_01-2023' and 'CREA Phenotypic hystorical assay\_01-2023 (Validation Report v4)'), and 'CRI-Phenotypic-Historical-Wheat' (containing another 'CRI-Phenotypic-Historical-Wheat' entry).

# • Data Integration between silos, From Phenotyping to Genotyping

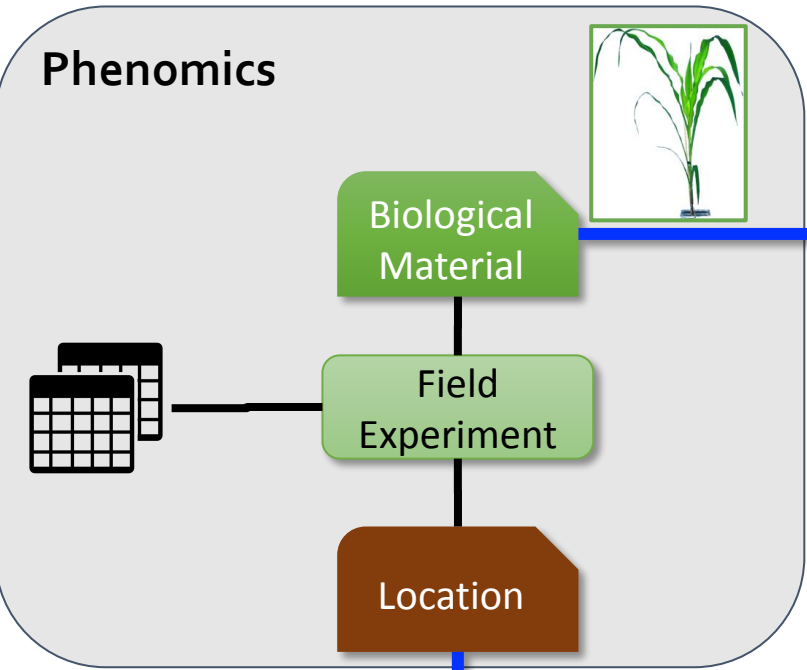


Identifying key resources/pivot objects

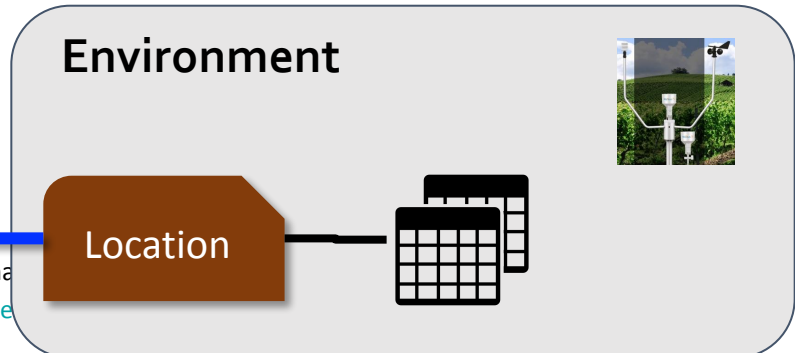
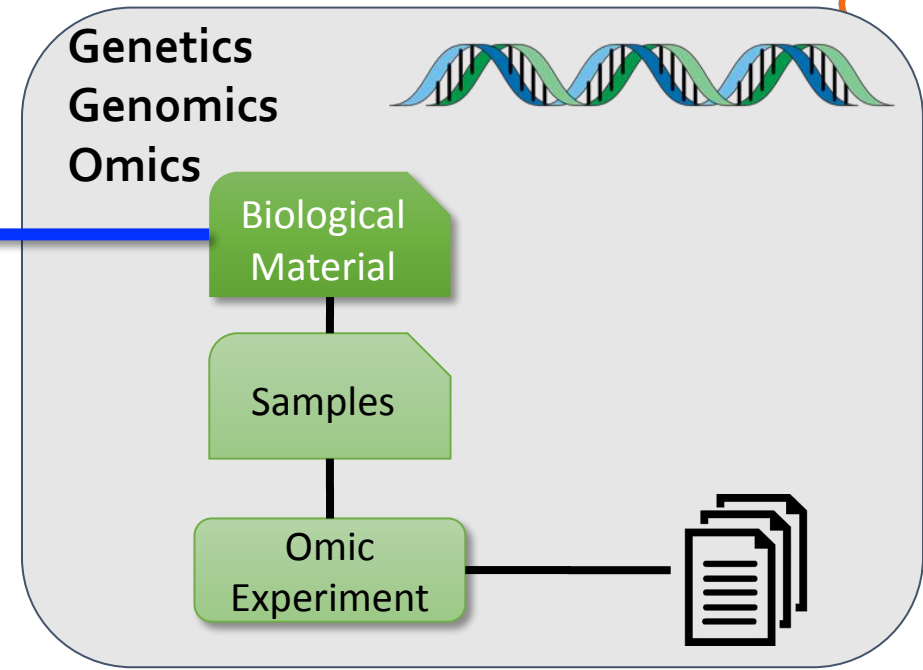




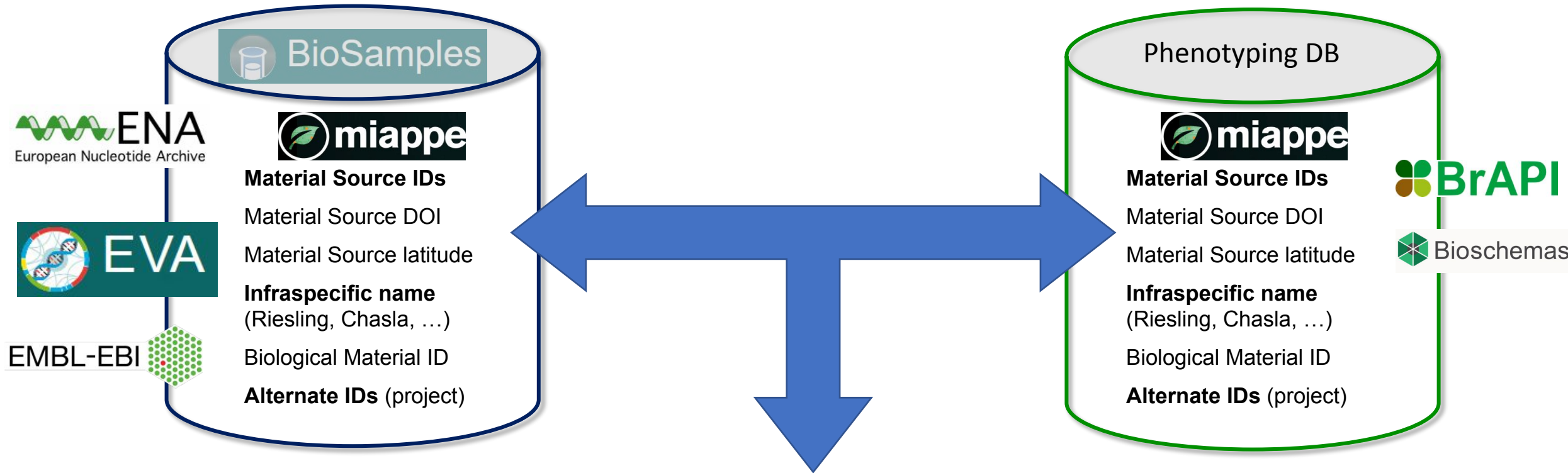
# • Data Integration between silos, From Phenotyping to Genotyping



Interoperability pivot  
Key shared resources



# Data Integration between silos, From Phenotyping to Genotyping



Community data discovery portals

URGI Data providers More... <https://urgi.versailles.inrae.fr/faidare/> elixir

### FAIR Data-finder for Agronomic REsearch

Search keywords

**Germplasm** Trait Reset all

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

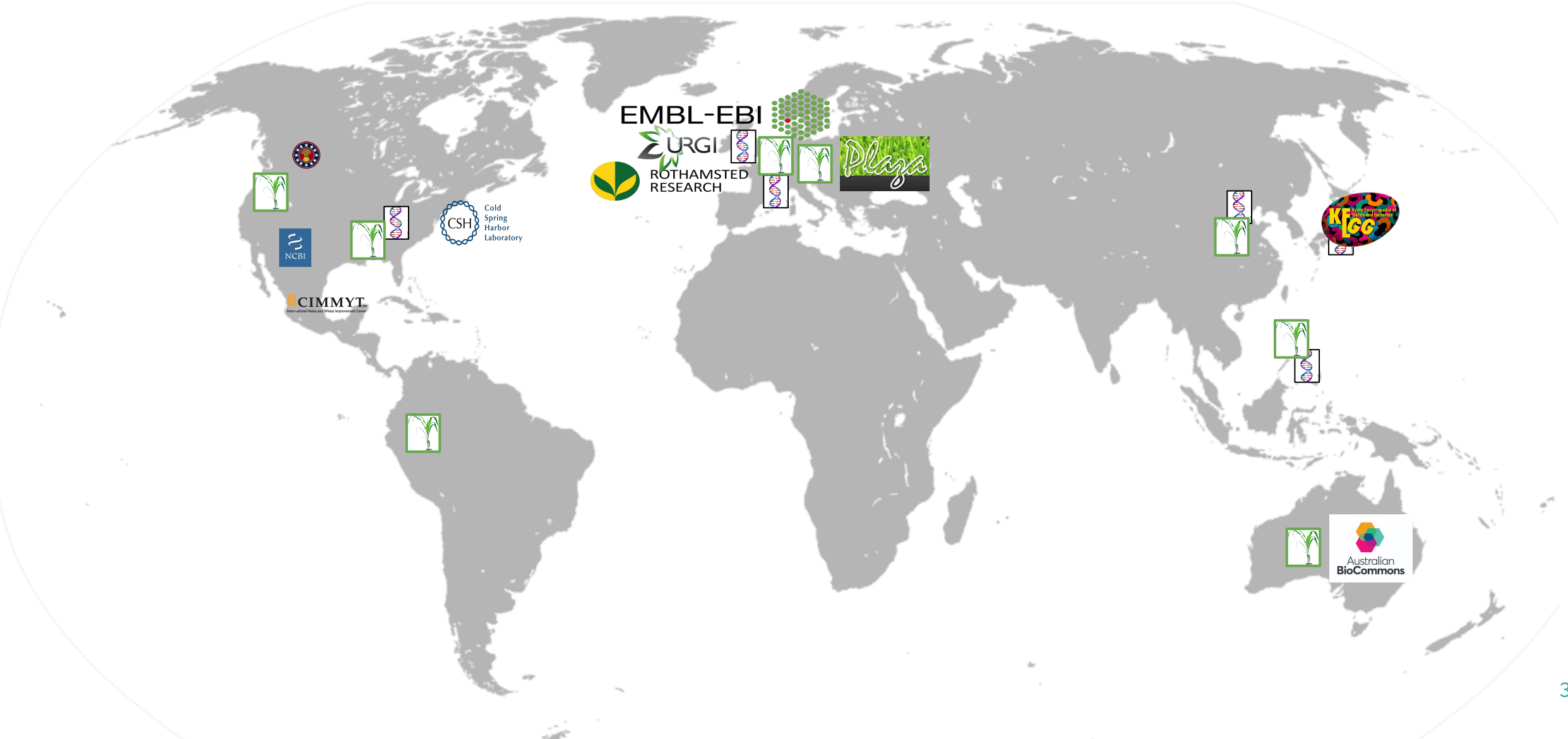
**Germplasm list** (panel, collection & population) Search germplasm lists

**Sources**

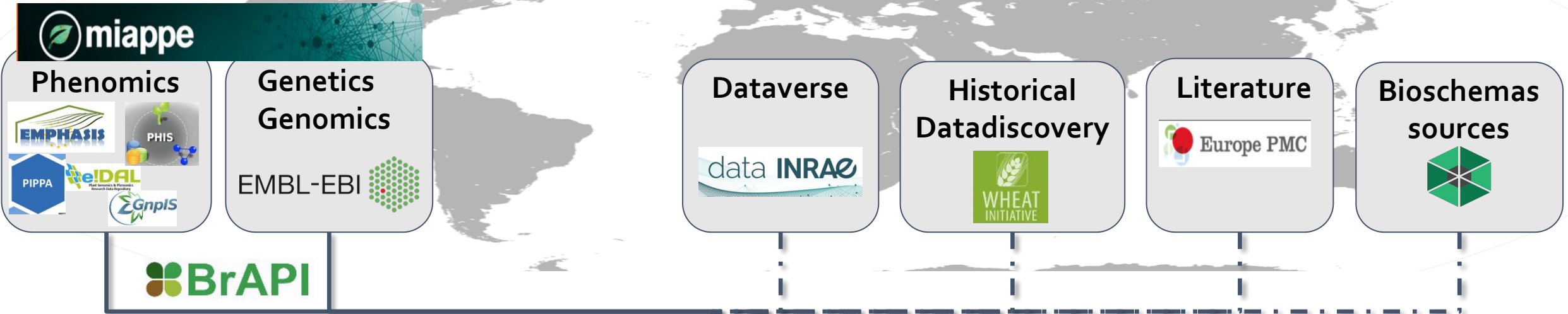
- URGI GnpIS (81,335)
- EBI European Nucleotide Archive (44,975)
- CIRAD TropGENE (722)
- VIB PIPPA (692)
- IBET BioData (67)
- IWGSC@GnpIS (18,814,632)
- Evoltree@GnpIS (5,354)
- OpenMinTeD@GnpIS (3,392)
- EBI Ensembl Plants

# • Global Data discovery portal

Dispersed data    Heterogenous data    Dedicated repositories & Archives



# FAIDARE: Global Data discovery portal



FAIDARE URGI More...

yield

Results 1 to 20 of 156

**Species (21)**  
Filter on Species...

**Data type**  
 Bibliography [151]  
 None [5]

**Ontology annotation (20)**

**10.3389/fpls.2018.00529** - OpenMinTeD@GnpIS  
 Bibliography **Triticum Triticum aestivum**  
 Global QTL Analysis Identifies Genomic Regions on Chromosomes 4A and 4B H...  
 Related Traits Across Different Environments in Wheat (Triticum aestivum L.). 20...  
 Genomic Regions on Chromosomes 4A and ... (expand)

**10.1186/s12864-019-6005-6** - OpenMinTeD@GnpIS  
 Bibliography **Triticum Triticum aestivum**  
 Genome-wide association study reveals new loci for **yield**-related traits in Sichu...  
 stripe rust stress. 2019 Genome-wide association study reveals new loci for **ye**...

Ontology variable selection

Filter English

- Woody Plant Ontology **Ontology**
  - Biochemical **Trait class**
  - Morphological **Trait class**
  - Other **Trait class**
  - Phenological **Trait class**
  - Budflush **Trait**
    - BF\_score\_BI: Broadleaves budflush scoring **Variable**
  - Budset date **Trait**
    - BS\_date: Budset date **Variable**

Identifier: CO\_357:1000009  
 Name: Budset date  
 Description: Assessment of the date when budset score will be reached for the first time  
 Entity: bud  
 Attribute: budset  
 Class: Phenological  
 Main abbreviation: BS\_date  
 Status: Standard for INRAE  
 Bud date protocol **Method**  
 Identifier: CO\_357:2000014  
 Name: Bud date protocol  
 Description: Estimated date from polynomial regression of a time series of budflush or budset scores  
 Class: Computation  
 Calendar day **Unit**  
 Identifier: CO\_357:3000043  
 Name: Calendar day  
 Data type: Date  
 Min: 0  
 Max: 0  
 Documentation: <https://urgi.versailles.inra...>  
 Context of use: Research-intensive characterization  
 Trial evaluation  
 Breeding criterion  
 Status: Standard for INRAE

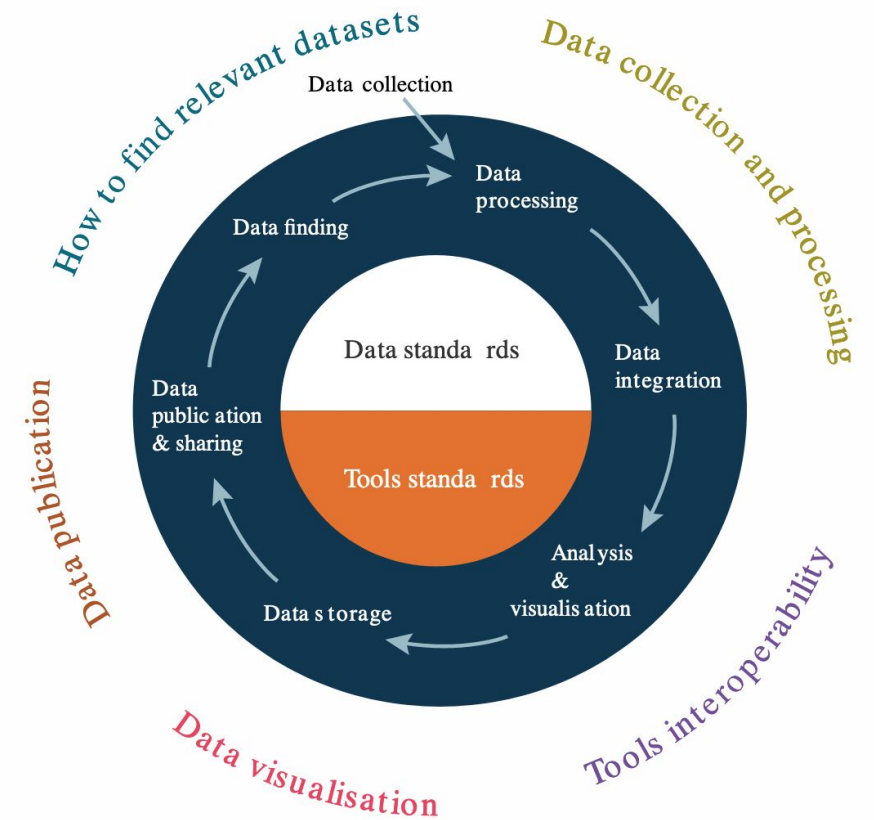
OK Cancel

Full text  
 +  
 Fine criteria  
 +  
Link back



# • Take Home Message

- Data management relies on a complex lifecycle
- Both:
  - Observation & Measures
  - Deriving and reducing data
  - Linking different datasets
- All steps must be defined
  - data management plan
- Not all step of data must be shared
- Raw and final data should be shared
- With sufficient provenance
- Community driven
- Data standard are easier to use
- Open science (policy): Publication and Findability are keys (Data tombs effect)
- Joint Training in 2024 on Pheno data management ? (Phenet: Sven Farhner, Jessica Lindvall)



# Aknowledgments

## Elixir Plant community & platforms

Beier S., Gruden C., Pommier C., Coppens F, Scholz U, Lange M., Contreras B., Adam Blondon AF, Faria D, Chavez I, Miguel C, Droedsbek B, Finkers R, Papoutsoglou E, Olster R, Ramsak Z, ...



## H2020 AGENT



N. Stein (IPK, coord), P. Kersey (RBGK), M. Alaux (INRAE), S. Weise (IPK), C. Pommier (INRAE), M. Lange (IPK), R. Finkers (WUR), J. Destin (INRAE)

## MIAPPE community



ELIXIR Plant Community, Krajewsky P, Cwiek H, Tardieu F, Usadel B, Arend D, Arnaud E, Junker A, King G, Laporte MA, Poorter H, Reif J, Rocca-Serra P, Sansone SA, Kersey P, And many more!



## Breeding API

Selby P, Mueller L, Robbins K, Backlund JE, ... , And many more!

## Crop Ontology



Arnaud E, Laporte MA, ...

## Emphasis



Tardieu F, Usadel B, Arend D, Junker A, Poorter H, Neveu P, Pierushka R, Shur U... And many more!



EMBL-EBI

