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

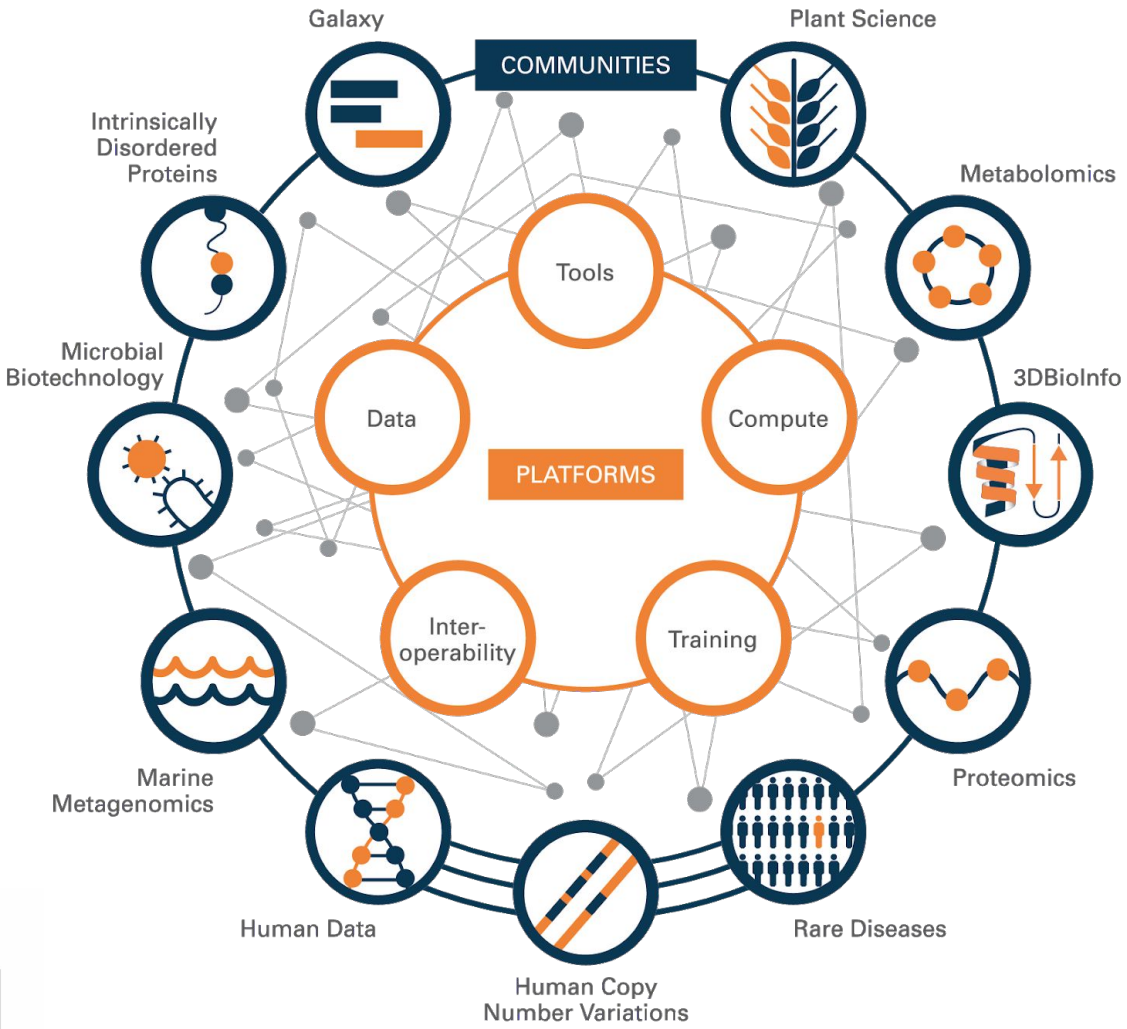
Solutions from ELIXIR and EMPHASIS European Infrastructure and beyond

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)

https://elixir-europe.org/communities/plant-sciences

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

elixir

• Open science through FAIR data principles

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

Findable

Ids
Index
Metadata
Description

Accessible

Open Protocole
Perenial Metadata

Interoperable

Semantics
Linked Data
Vocabularies

Reusable

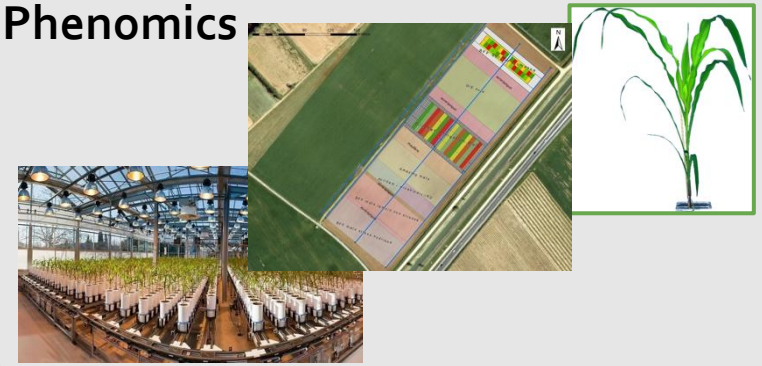
License
Well described
Provenance (origin, process, methodology)
Standards

Sustainable data access over decades

• PLANT use case

• Environment / Phenome / Genomic / *omic / Genetic

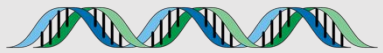
Phenomics

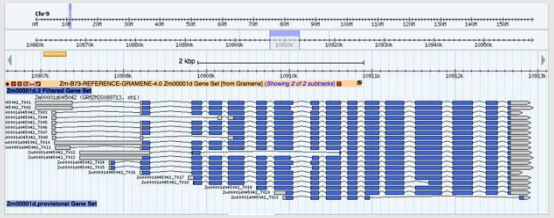


Dispersed
Heterogenous
Getting Standardized

Findable Accessible Interoperable Reusable

Genetics
Genomics
Omics





Mostly centralized
Homogenous data
Heterogenous metadata

Plant Breeding
Genetic variations by Traits

Climate Change Studie
Genotype by Environment

Findable Accessible Interoperable Reusable

Dispersed
Heterogenous

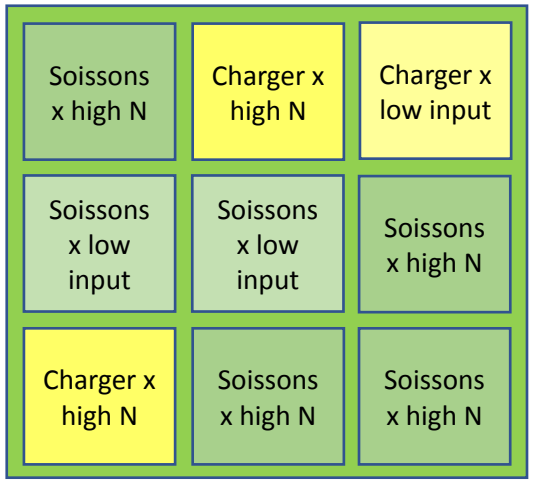
Environment

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

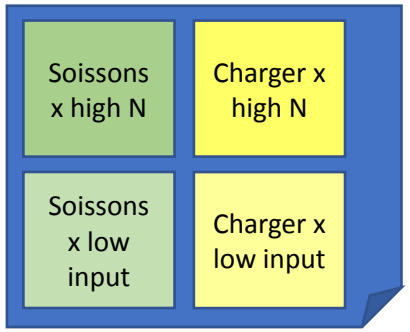
Phenotyping data life cycle

« Raw » data, pheno/env measures, variables



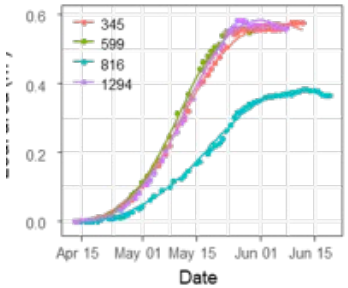
Derivation, Reduction

« 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	71.739999999999995	01/01/2004	305.7
661300585	Cavallermaggiore	2005			
661300468	Headley	2004	45.27	01/01/2004	
661300468	Headley	2005			
661300469	Headley	2004	70.930000000000007	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

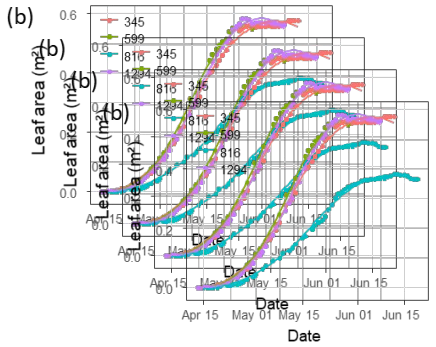
- **VARIABLES**
- Plant/microplot level
- Traceability
- Raw measures
- Data Cleaning
- Platform IS (Emphasis IS, PHIS, …)
- Analysis Reproducibility
- Provenance

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

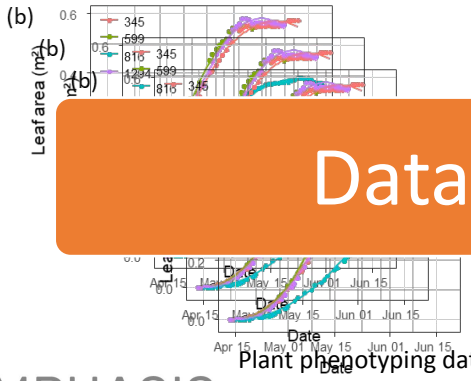
- **VARIABLES**
- Plant/microplot level
- Traceability
- Raw measures
- Data Cleaning
- Platform IS (Emphasis IS, PHIS, …)
- Analysis Reproducibility
- Provenance

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		
Charger	low input	1
Charger	high N	2

Variety charger

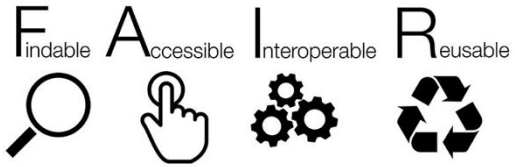
intensiv cultural practice

Knowledge

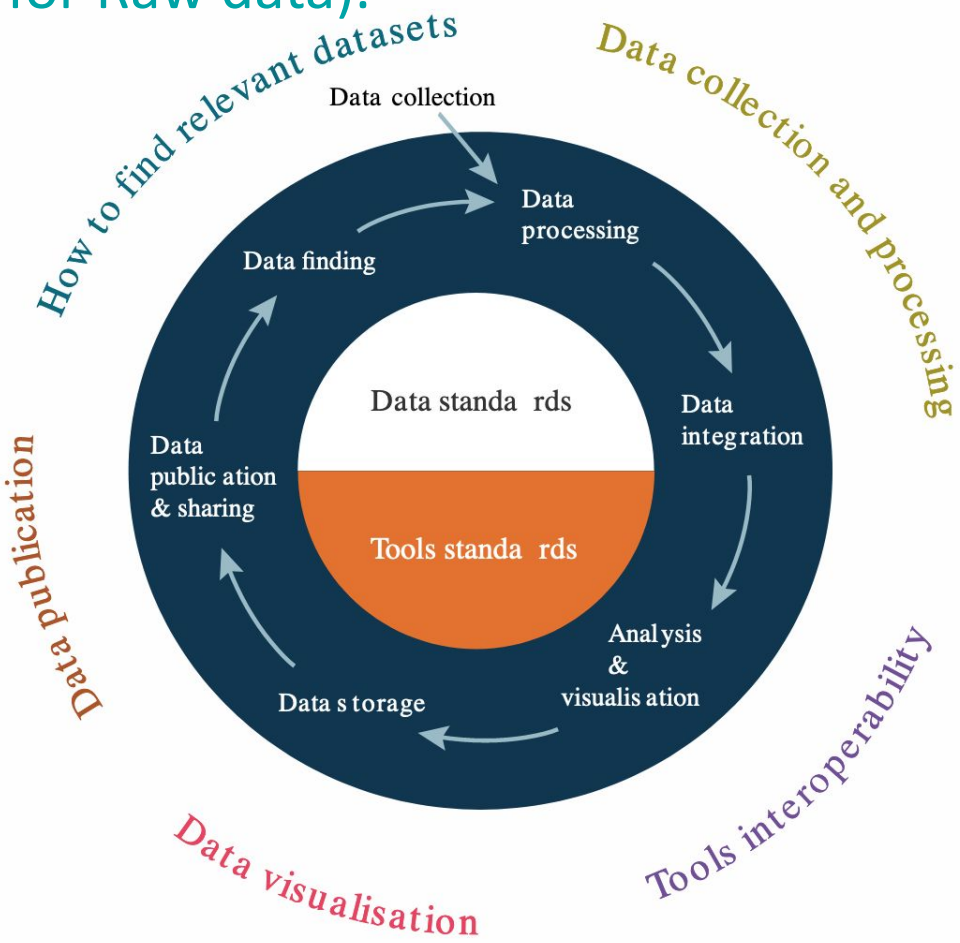
• Plant Genetic variation

- 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

Phenotype 1 = measurement on a genotype in an environment-GPS1-time1

Phenotype 2 = measurement on a genotype in an environment-GPS2-time2

Genotype = observed marker's alleles on a genotype

Climate 1 = climatic data at GPS1-time1

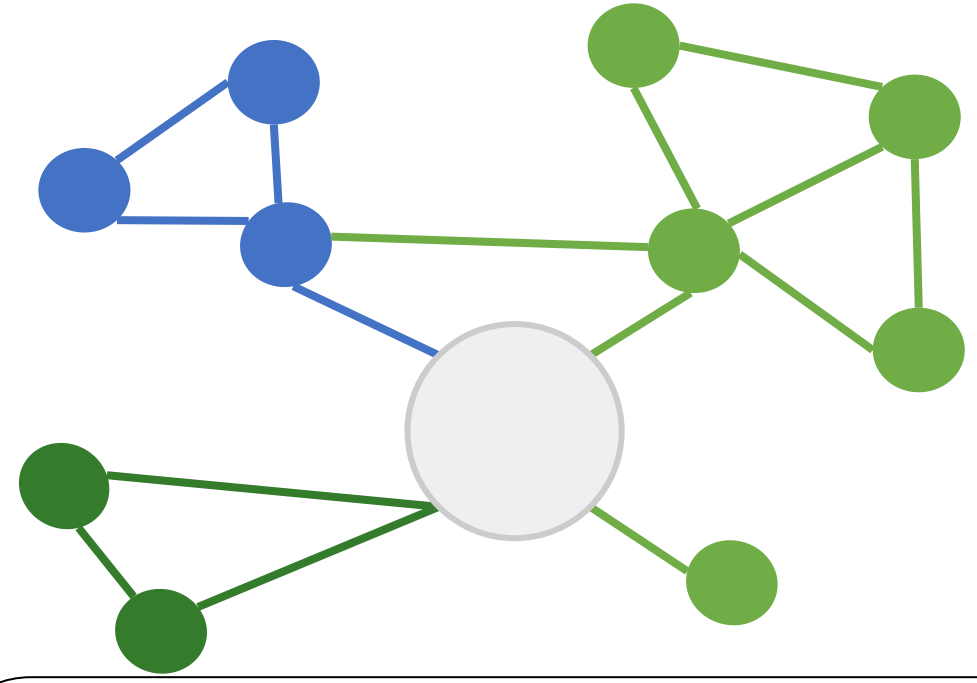
- To enable knowledge discovery: metadata about the experiment, controlled vocabularies, ontologies

- **PLANT DATA STANDARDS : WHO**




• Interoperability in International network

National Networks



European Networks

Global Networks

International data standards

BrAPI
Web services

RDA RESEARCH DATA ALLIANCE

miappe
Minimal information

Crop Ontology for agricultural data
Controlled vocabularies
Trait dictionaries

MCPD

• Sharing standards: standards registries

FAIRsharing.org
standards, databases, policies

Search all of FAIRsharing | Standards | Databases | Policies | Collections | Add/Claim Content | Stats | Lo

How to cite this record FAIRsharing.org: GnpIS; Genetic and Genomic Information System; DOI: 10.25504/FAIRsharing.dw22y3; Last edited: May 8, 2018, 9: 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

- **Community driven recomendations and registries**

- WheatIS: http://wheatis.org/DataStandards.php
- RDM Toolkit
  - **https://rdmkit.elixir-europe.org/**

- Community story

F1000Research
Open for Science
Search
BROWSE GATEWAYS & COLLECTIONS HOW TO PUBLISH ABOUT

RDMkit | Data management | About | Contribute | GitHub | Search

Data management

- Data life cycle
- Your role
- Your domain
- Your tasks
- Tool assembly
- National resources

- All tools and resources
- All training resources

Are you working with data in the Life Sciences? Do you fee overwhelmed when you think about Research Data Mana

The ELIXIR Research Data Management Kit (RDMkit) is an online guide contle management practices applicable to research projects from the beginning che[5], Laurel Cooper [6], Hanna Ćwiek-Kupczyńska[7], Robert P. Davey [8], Developed and managed by people who work every day with life science d an Fulss[9], Clement Jonquet [10,11], Marie-Angélique Laporte[3], Pierre Larmande [12,13], has guidelines, information, and pointers to help you with problems throu nier [2], Vassilis Protonotarios [14], Carmen Reverte [15], Rosemary Shrestha[9], life cycle. RDMkit supports FAIR data — Findable, Accessible, Interoperable rats[16], Aravind Venkatesan [12], Alex Whan[17], Hadi Quesneville [2] by-design, from the first steps of data management planning to the final st data in public archives.

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

Data life cycle

Check for updates

RTICLE

Developing data interoperability using standards: at community use case [version 2; referees: 2 ved]

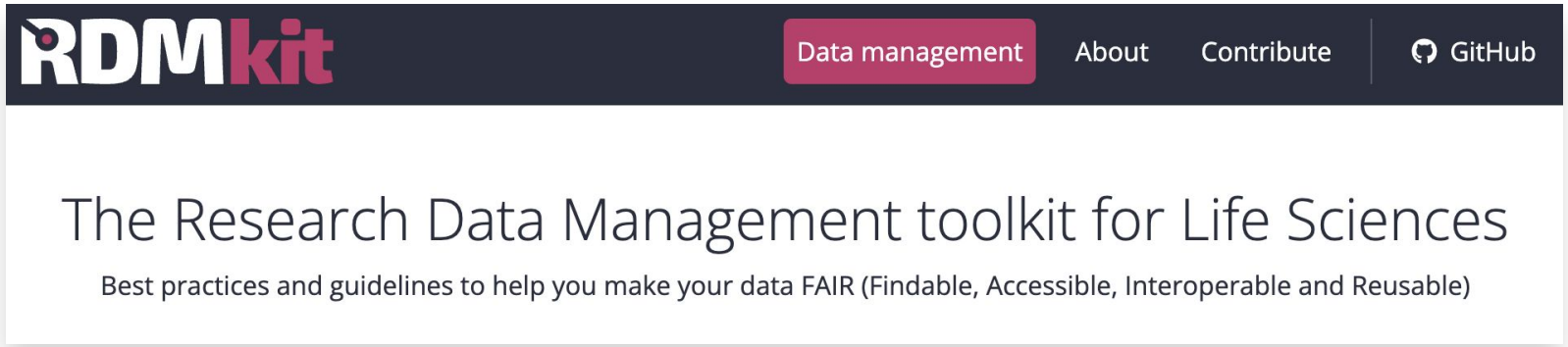
le Yeumo[1], Michael Alaux [2], Elizabeth Arnaud[3], Sophie Aubin[1], Ute Baumann[4], che[5], Laurel Cooper [6], Hanna Ćwiek-Kupczyńska[7], Robert P. Davey [8], an Fulss[9], Clement Jonquet [10,11], Marie-Angélique Laporte[3], Pierre Larmande [12,13], nier [2], Vassilis Protonotarios [14], Carmen Reverte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

letails

This article is included in the Global Open Data for Agriculture and Nutrition gateway.

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

179

Contributors
The force behind RDMkit

514

Tools & resources
Explained in the context of real world problems

122

Pages
Helping you with data management

- 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

from March 2021 until September 2023

Plant Pages including PGR

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

- **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

Structure

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

Persistent Unique Identifiers
URI, gene ID, accessions ID, Trait ID, DOI,…

Technical

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

• Semantic Standard: Ontologies

- Annotating one object
  - Protein, gene
  - Plant, plant anatomy, …

- Atomic concept
  - protein function
  - cellular localization
  - …

Gene Product	Symbol	Qualifier	GO Term	Evidence	Reference	With / From	Taxon
UniProtKB:P56761	psbD	is_active_in	GO:0009535 (C) chloroplast thylakoid membrane	ECO:0000318 IBA	PMID:21873635	PANTHER:PTN002108145 more...	3702 Arabidopsis thaliana
UniProtKB:P56761	psbD	part_of	GO:0009523 (C) photosystem II	ECO:0000318 IBA	PMID:21873635	PANTHER:PTN002108145 more...	3702 Arabidopsis thaliana
UniProtKB:P56761	psbD	enables	GO:0005515 (F) protein binding	ECO:0000353 IPI	PMID:25846821	UniProtKB:Q9FL44	3702 Arabidopsis thaliana
UniProtKB:P56761	psbD	involved_in	GO:0019684 (P) photosynthesis, light reaction	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

AgroPortal LIRMM

Ontology Lookup Service

Crop Ontology for agricultural data

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

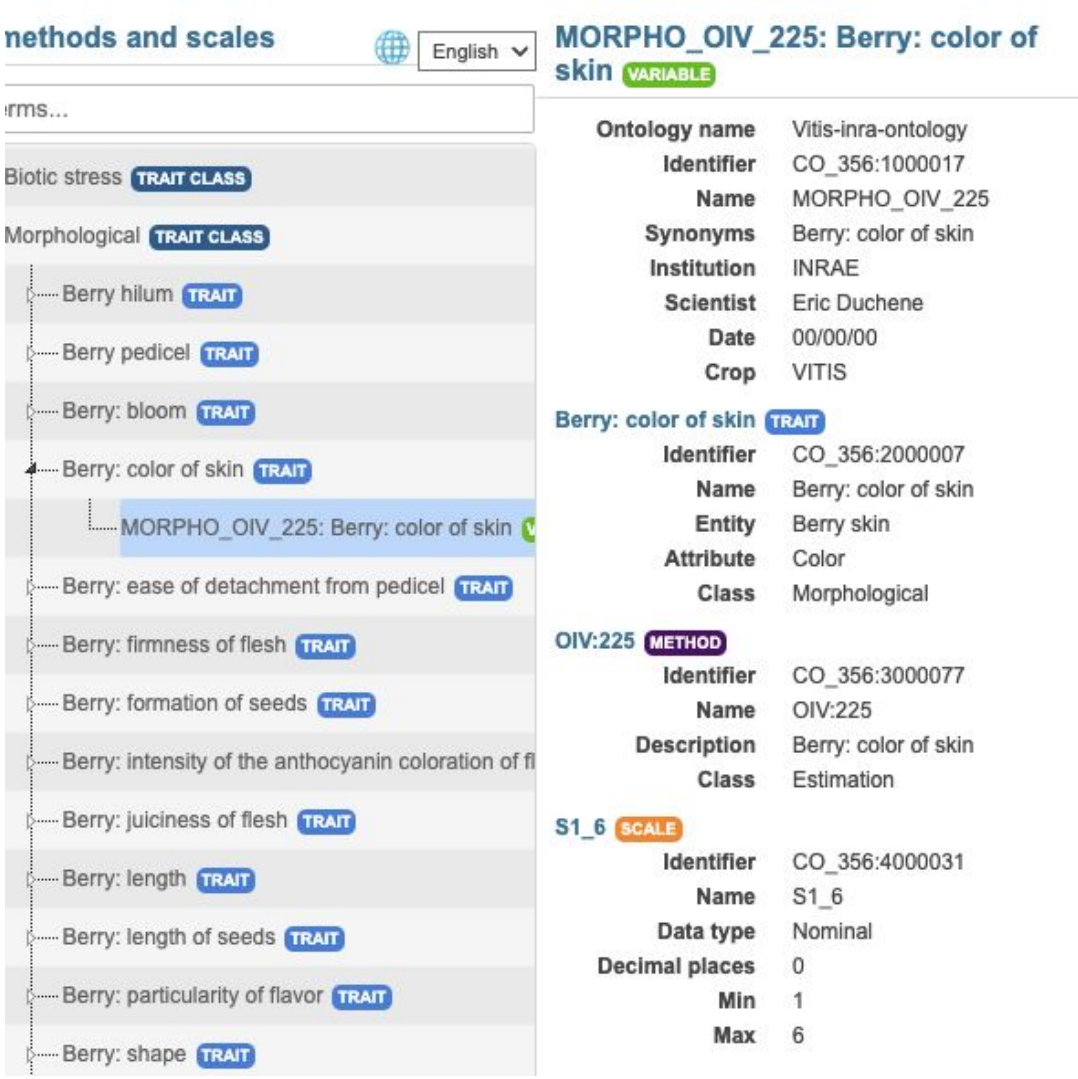
M5: Highest pixel corresponding to plant

U3: pixel

Slide from L. Cabrera-Bosquet

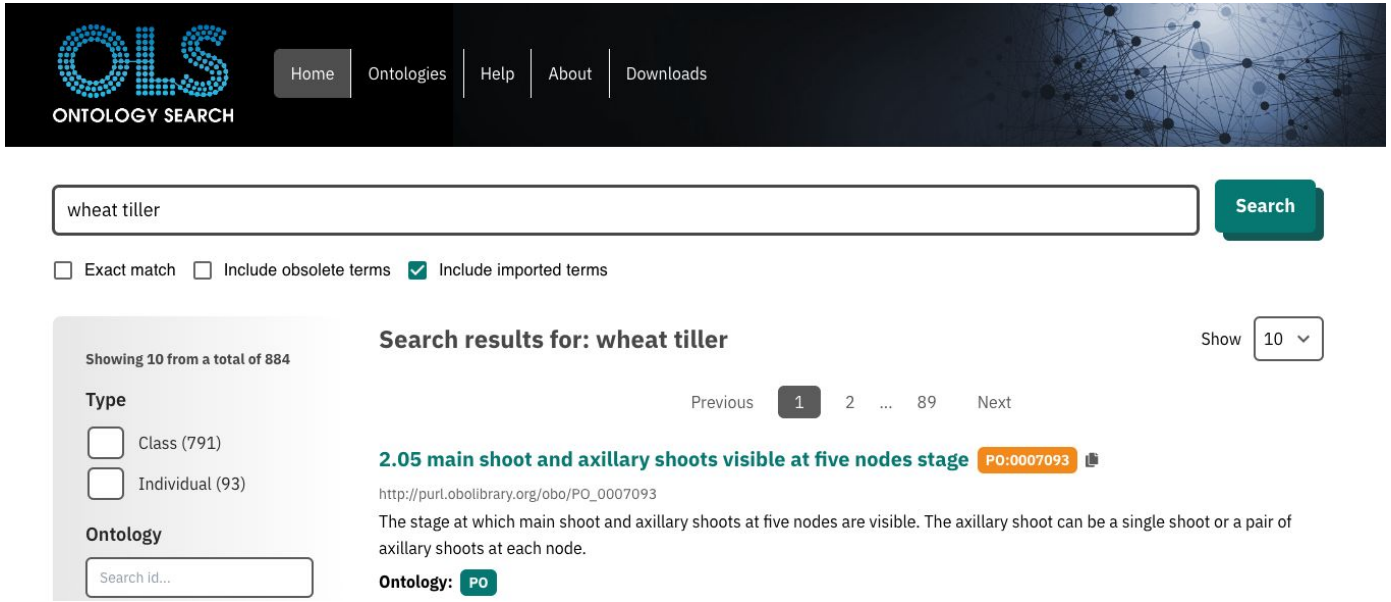
Ontologies, variables, descriptors

- Phenotyping variables
  - Crop Ontology
    - Collection of species specific ontologies
    - https://www.cropontology.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

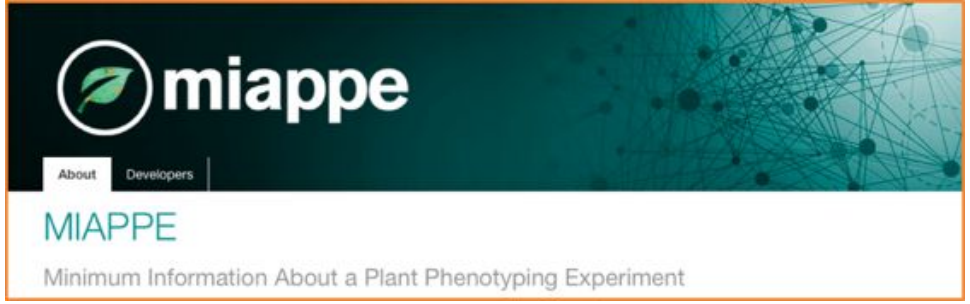


Ontologies, variables, descriptors

- Phenotyping variables
  - Crop Ontology
    - Collection of species specific ontologies
    - https://www.cropontology.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



• Phenotype Structure Standard



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

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

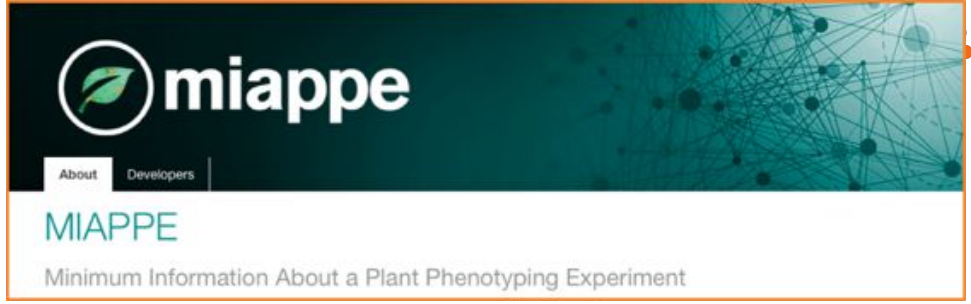
	MIAPPE				
line #	MIAPPE Check list	Definition	Example	Format	Cardinality
DM-1	**Investigation**	Investigations are research programmes with defined aims. They can exist at various scales (for example, they could encompass a grant-funded programme of work, the various components comprising a peer-reviewed publication, or a single experiment).			1 per MIAPPE submission
DM-2	**Investigation unique ID**	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.	EBI:12345678	Unique identifier	0-1
	Investigation title	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

line#	Environment			
ENV-1	Non-exhaustive list of Environment Parameters.			
ENV-2	**Environment parameters**	**Definition**	**Example**	**Format**
ENV-3	**Growth facility**			
ENV-4	**Air temperature**	List of hourly air temperature throughout the experiment.	22 °C	Numeric
ENV-5	**Organ temperature**	List of hourly organ temperatures throughout the experiment	18 °C	Numeric

line #	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	**Seasonal environment**	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	**Air treatment regime**	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	**Soil temperature regime**	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'

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**

• Phenotype Structure Standard



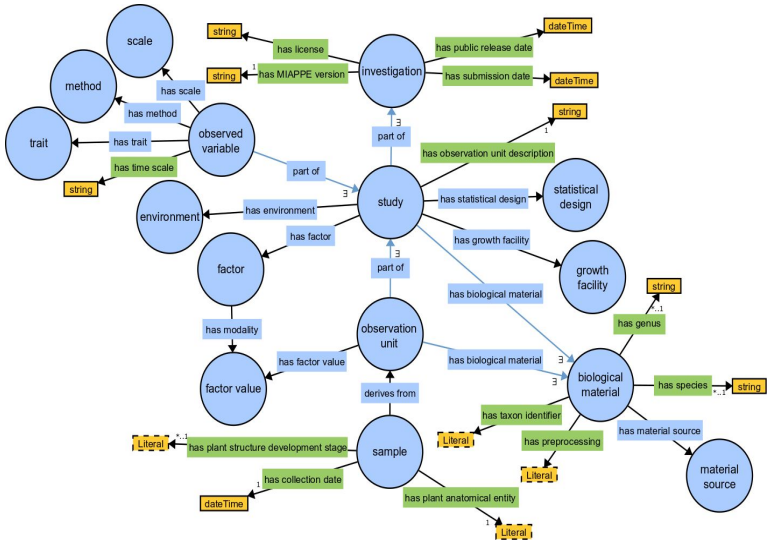
Minimum Information for Biological and Biomedical Investigations

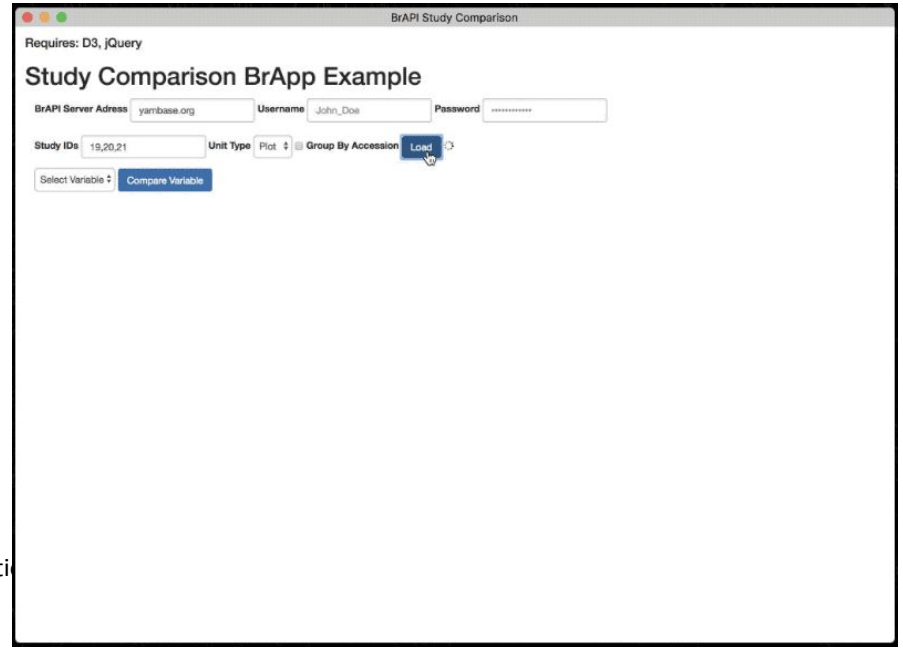
A collection of the historical MIBBI foundry reporting guidelines. The minimum information standard is a set of guidelines for reporting data derived by relevant methods in biosciences. If followed, it ensures that the data can be easily verified, analysed and clearly

- 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

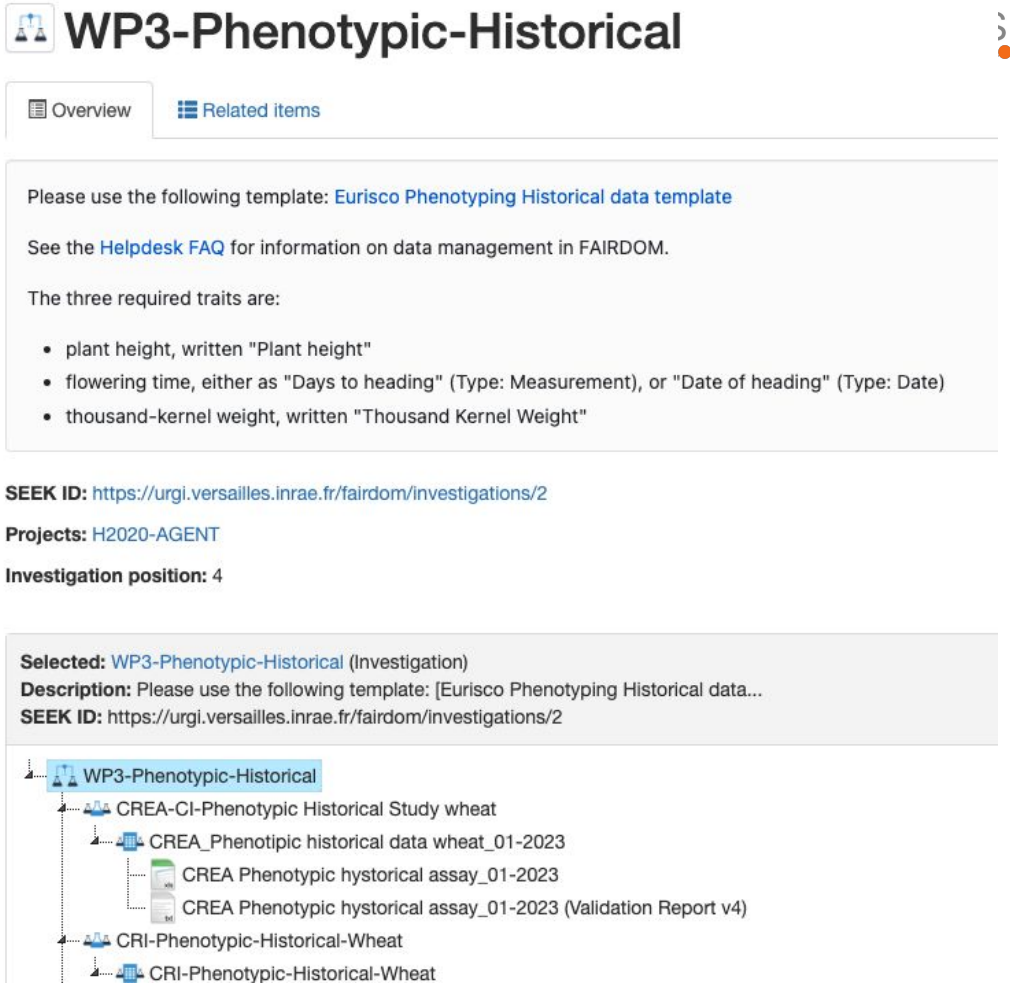
- Ontology, OWL Implementation
  - https://github.com/MIAPPE/MIAPPE-ontology
  - http://agroportal.lirmm.fr/ontologies/PPEO
  - Data model representation
  - Formal concepts and constraints
- File Archive
  - ISA Tab: data + metadata
  - RO Crate studies
- Web Services
  - Breeding API
  - International collaboration
  - Standard Open Web Service API
  - Information Exchange, Main target: Breeding
  - Excellence in Breeding platform (CGIAR, Peter Selby)



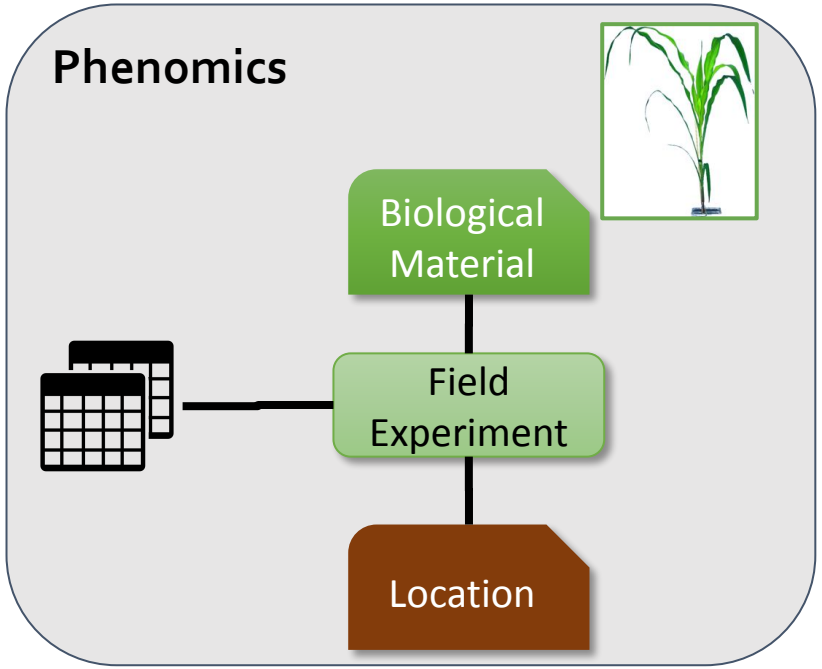


Data management tools

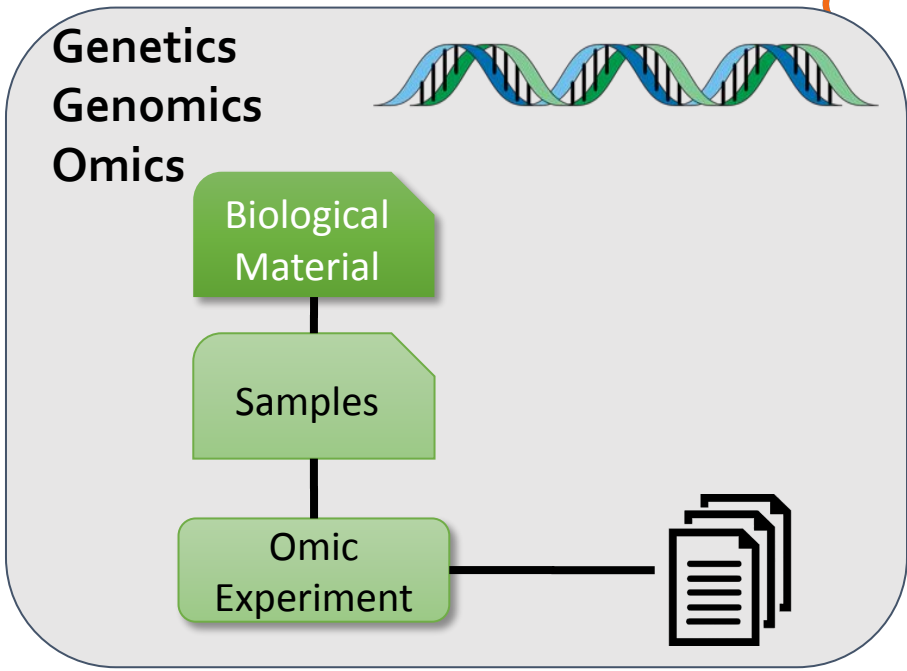
- Simple phenotyping network
  - Fairdom Seek (AGENT H2020)
  - Simple shared space for data, files SOP, etc..
  - Avoid “sent by email”
  - **Usable beyond Phenotyping**
- Fully instrumented Phenotype
  - Emphasis solutions (PIPPA, PHIS, IPK, …)
    - https://emphasis.plant-phenotyping.eu/services/emphasis-pilots/data-services

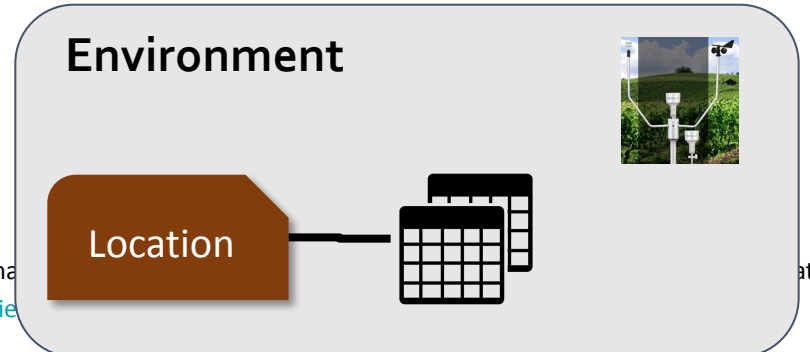


•Data Integration between silos, From Phenotyping to Genotyping

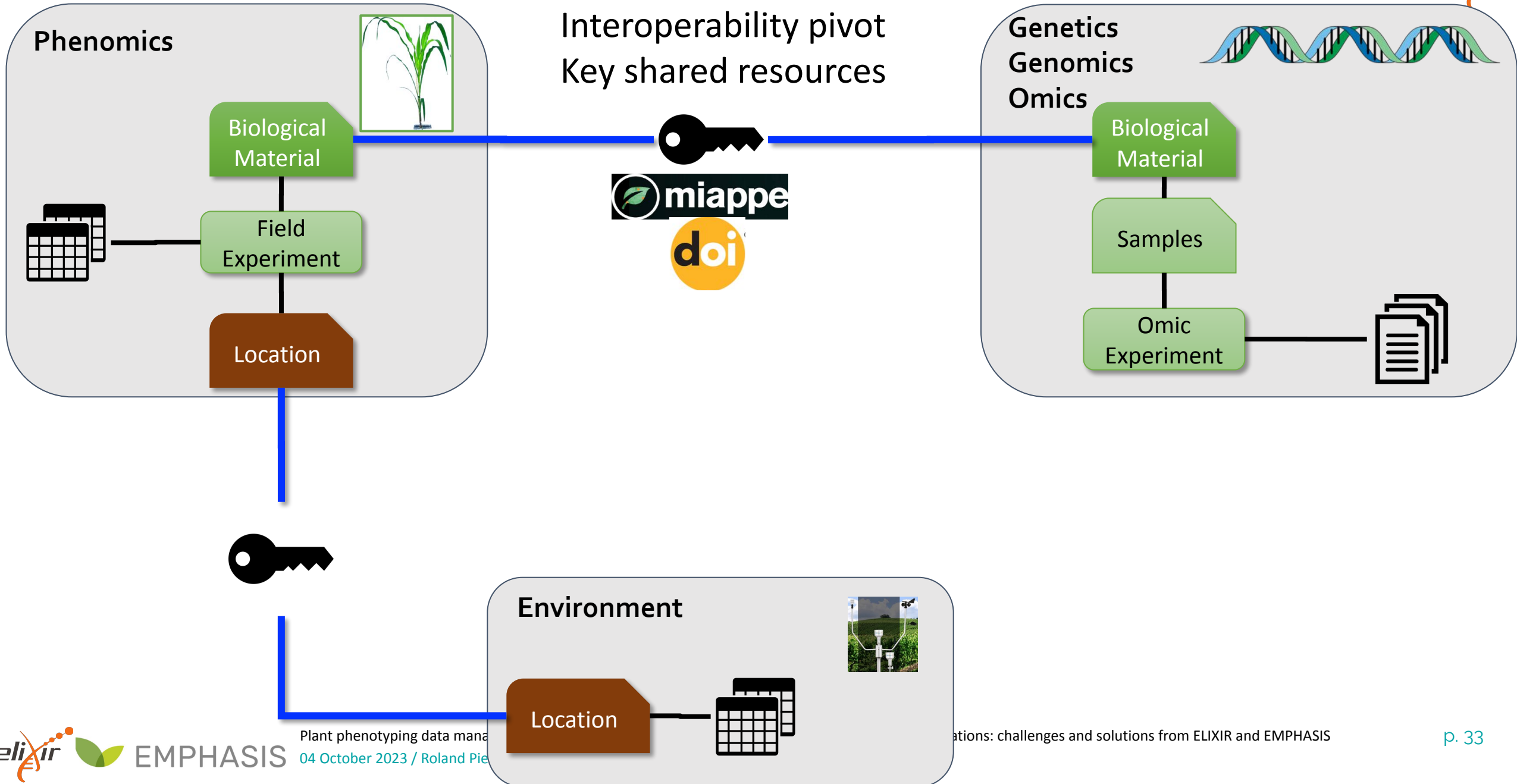


Identifying key resources/pivot objects

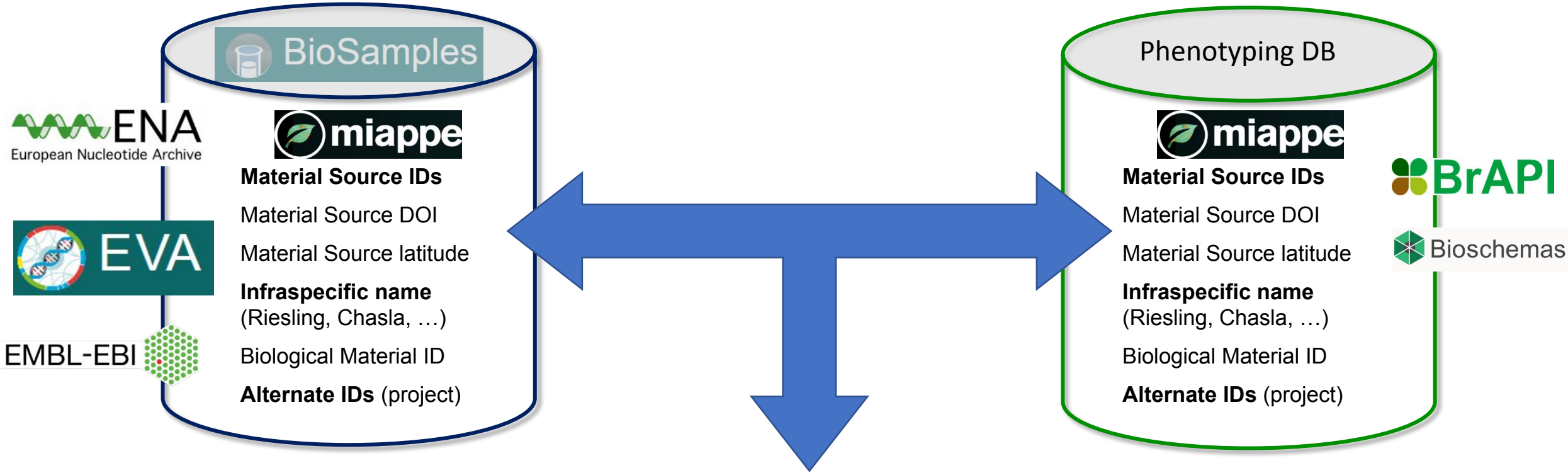




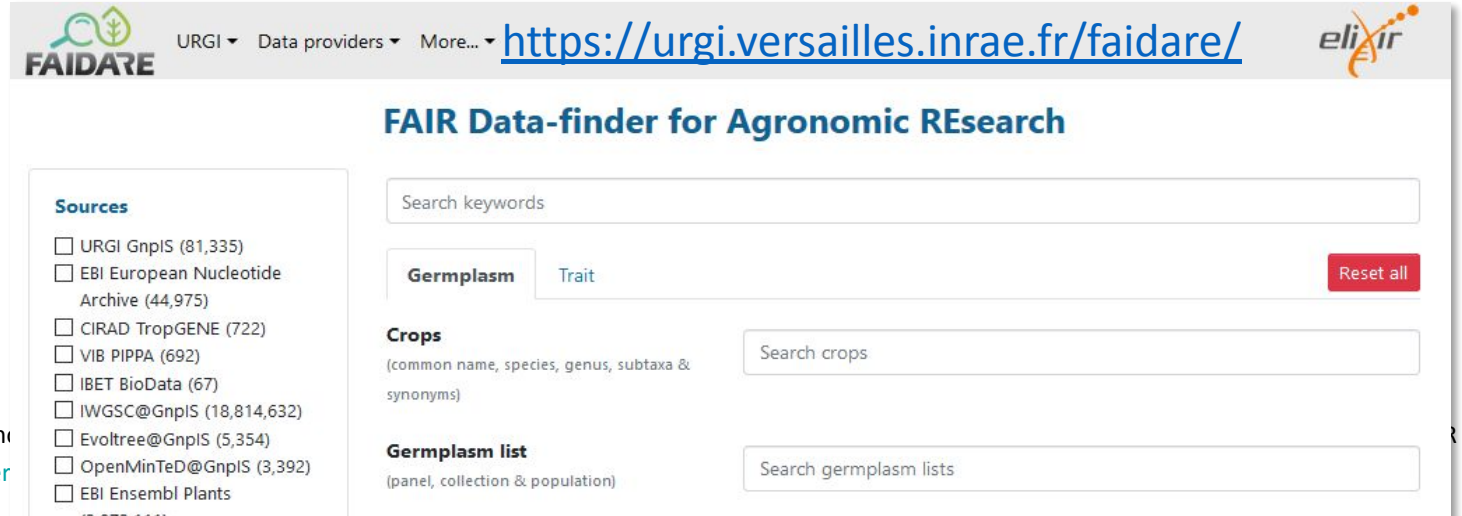
•Data Integration between silos, From Phenotyping to Genotyping



•Data Integration between silos, From Phenotyping to Genotyping

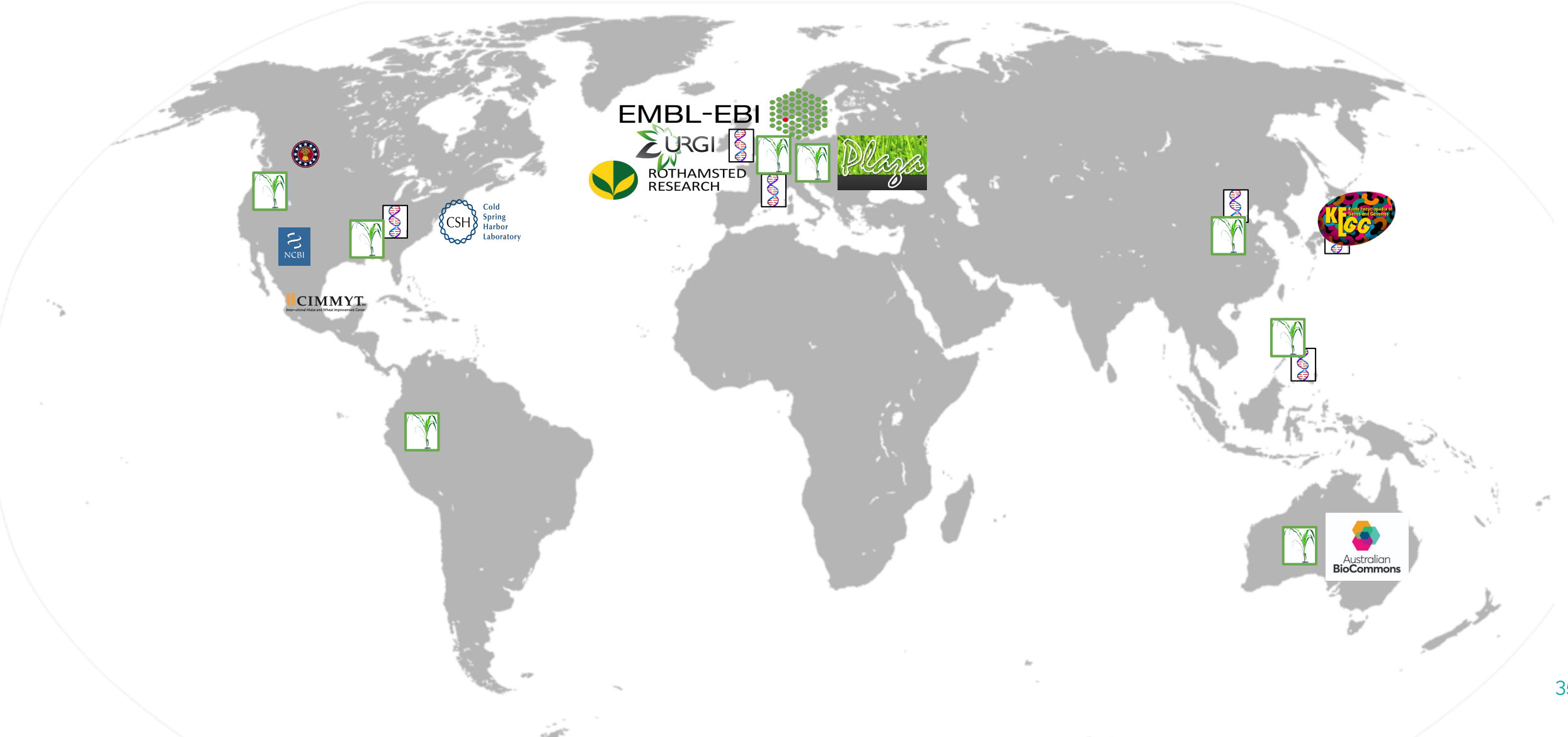


Community data discovery portals



Global Data discovery portal

Dispersed data Heterogenous data Dedicated repositories & Archives



• FAIDARE: Global Data discovery portal

FAIDARE

miappe

Phenomics

EMPHASIS, PHIS, PIPPA, eIDAL, GnpIS

Genetics Genomics

EMBL-EBI

Dataverse

data INRAE

Historical Datadiscovery

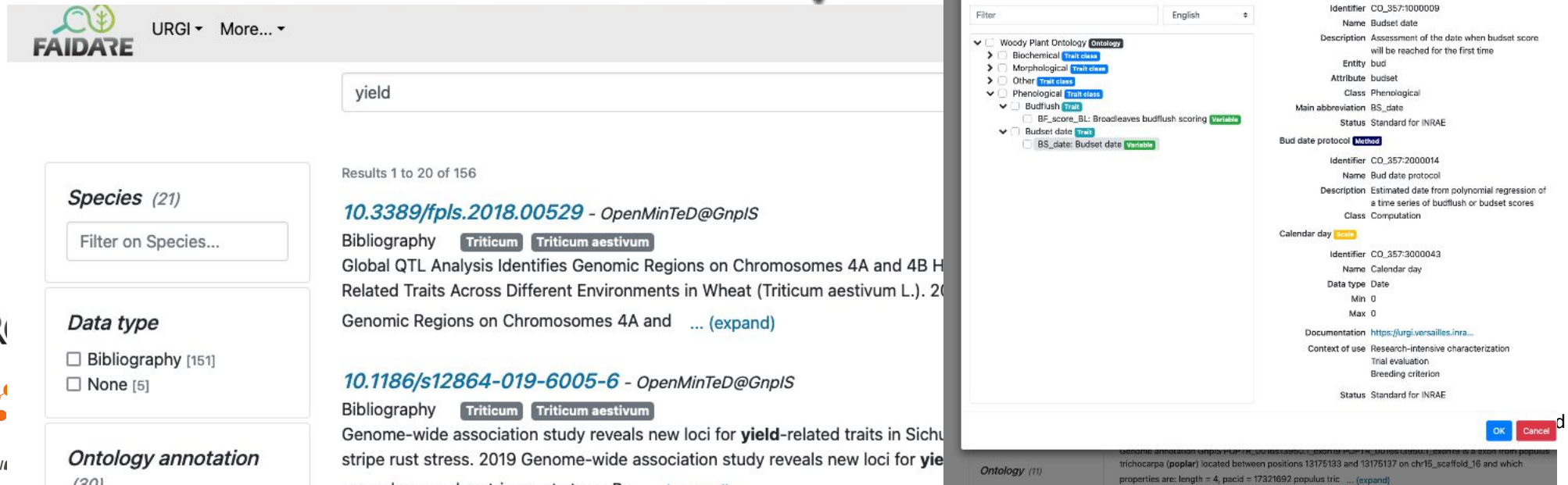
WHEAT INITIATIVE

Literature

Europe PMC

Bioschemas sources

BrAPI



Full text

+

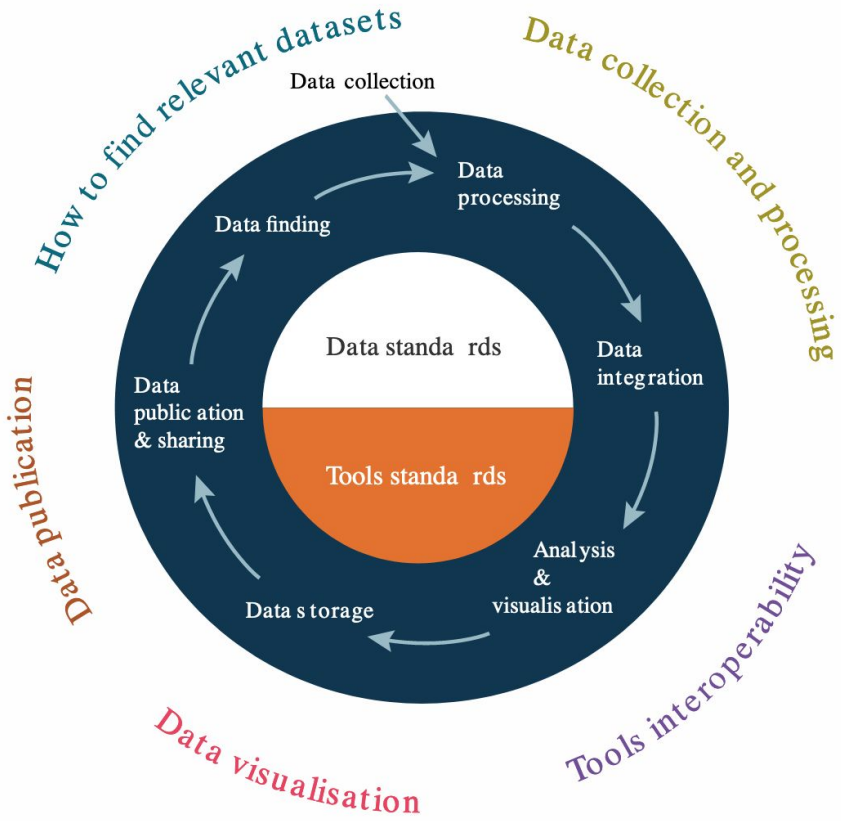
Fine criteria

+

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!

Crop Ontology

Arnaud E, Laporte MA, …

Breeding API

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

Emphasis

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