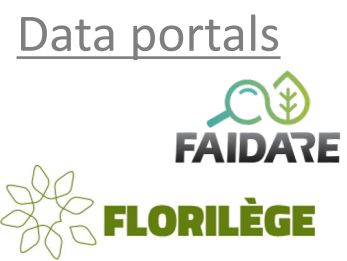
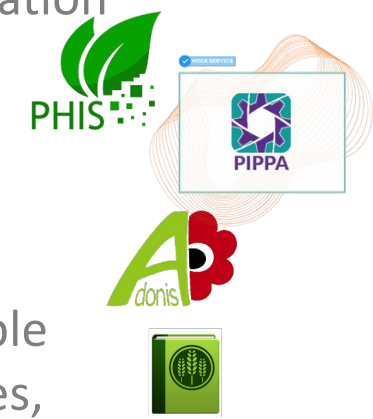


➤ FAIR Plant Phenomics Data Management Tools and Guidelines

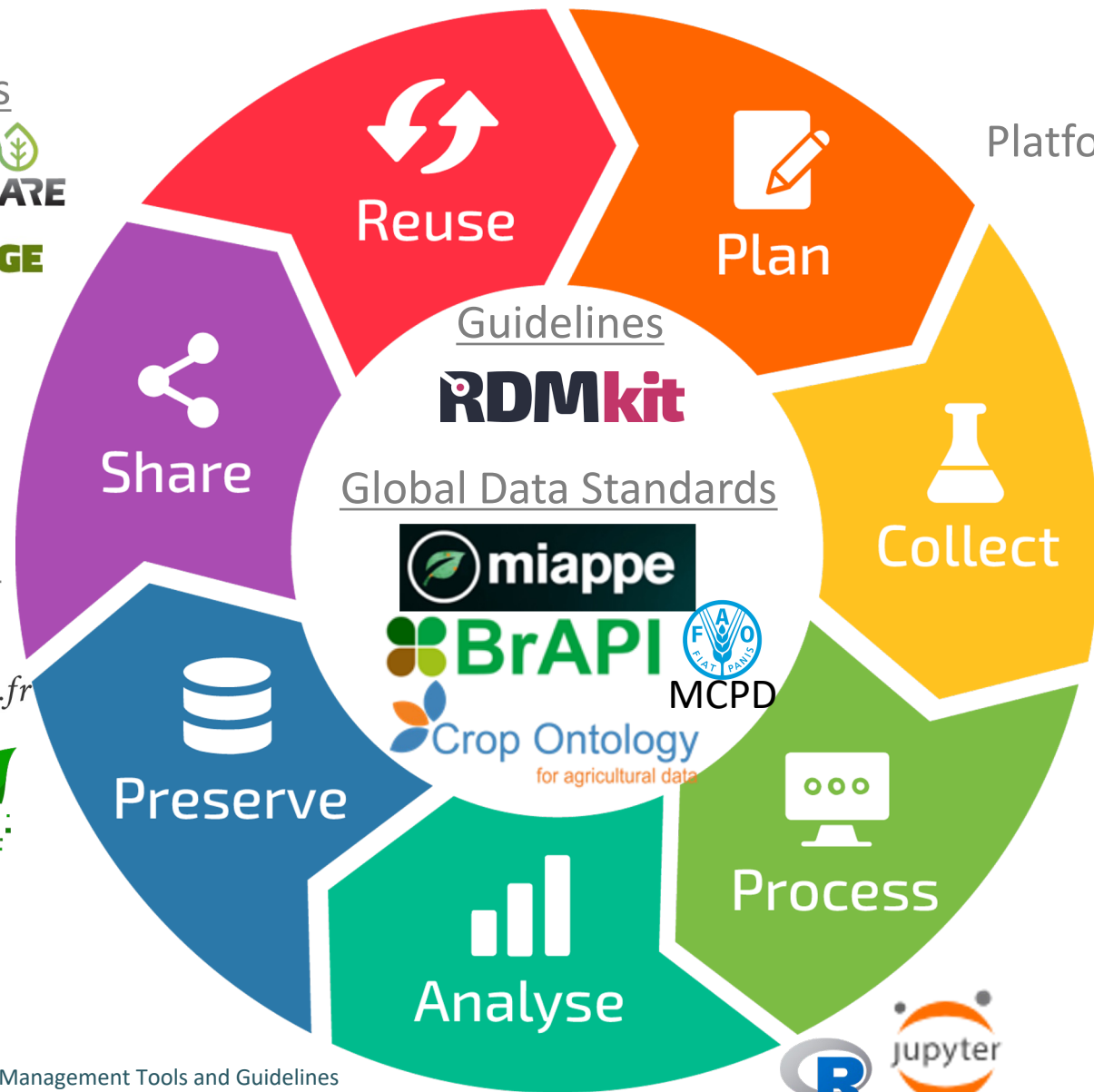
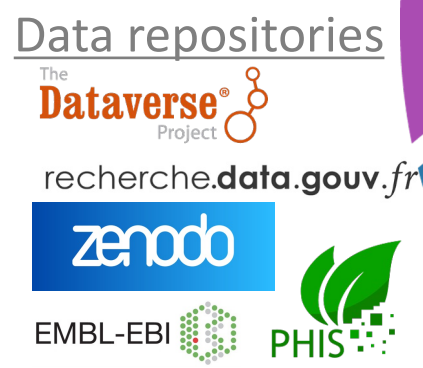
Some current practices in Elixir and Emphasis European Infrastructures



Platform Information Systems

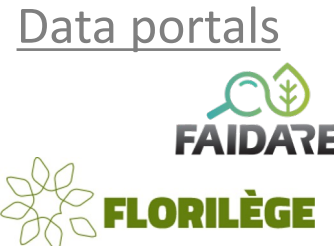


Portable devices, sensors, ...



Scripts and Workflows

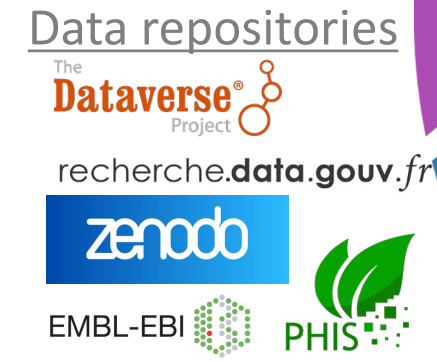
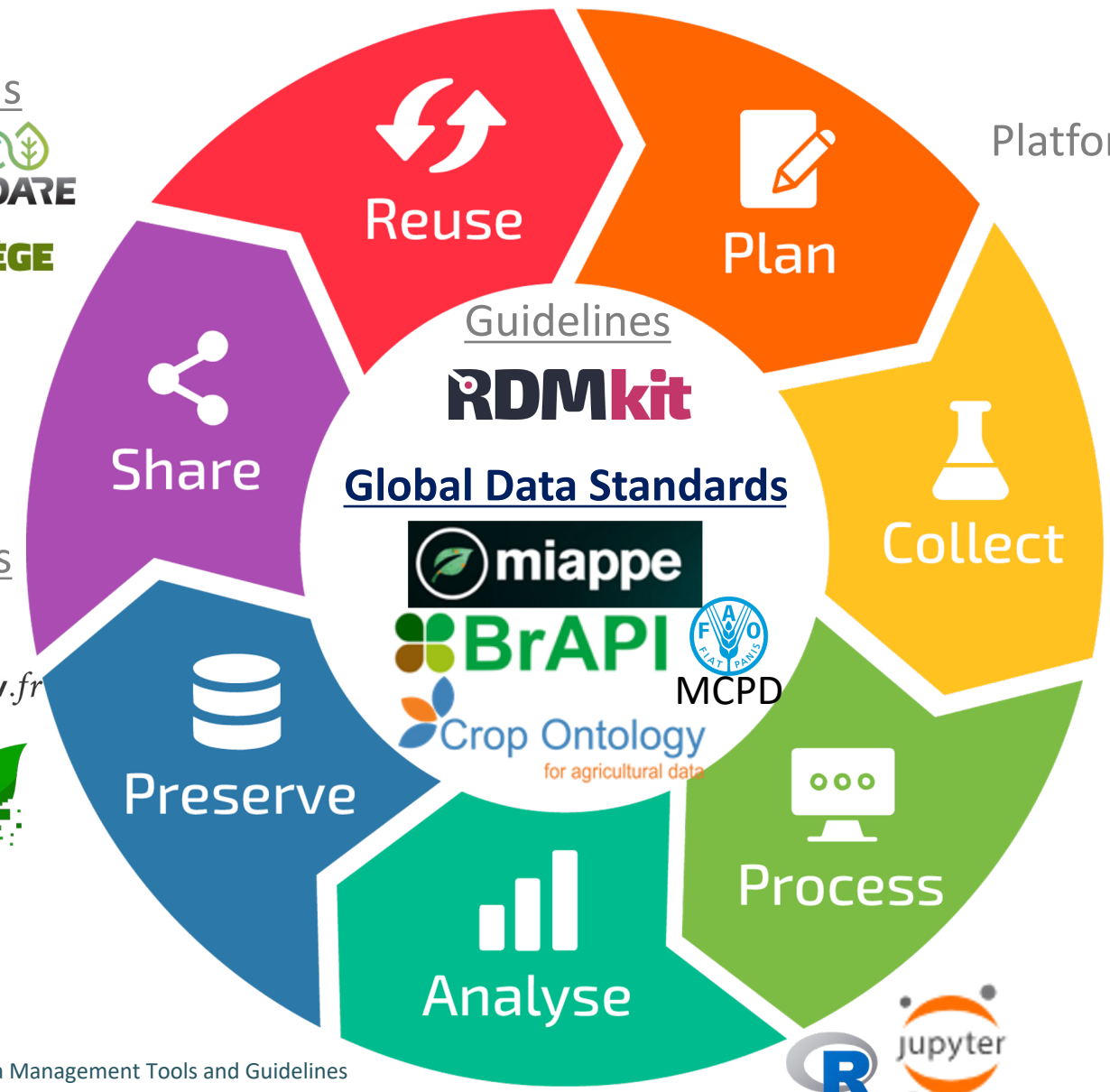




Platform Information Systems



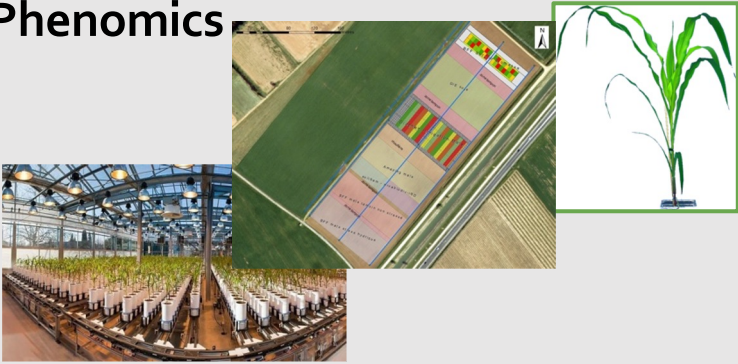
Portable devices, sensors, ...



Scripts and Workflows



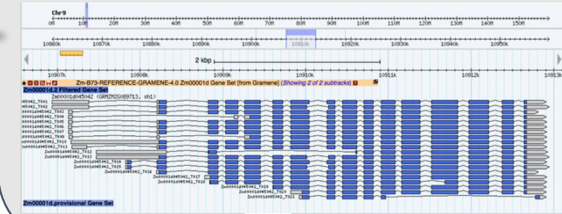
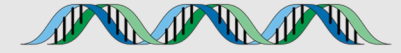
Phenomics



F_{indable} A_{ccessible} I_{nteroperable} R_{eusable}



Genetics Genomics Omics



Dispersed (no central repo)
Heterogenous
Getting Standardized

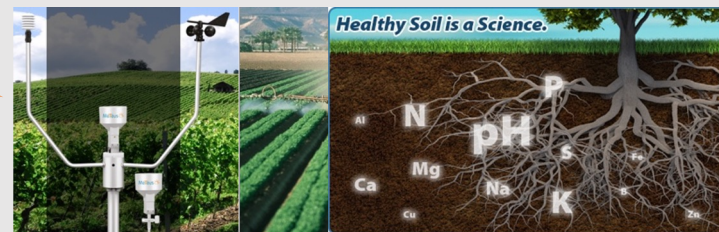
Plant Breeding
Genetic variations by Traits

Climate Change Studies
Genotype by Environment

Mostly centralized
Homogenous data
Heterogenous metadata

F_{indable} A_{ccessible} I_{nteroperable} R_{eusable}

Environment



Dispersed
Heterogenous

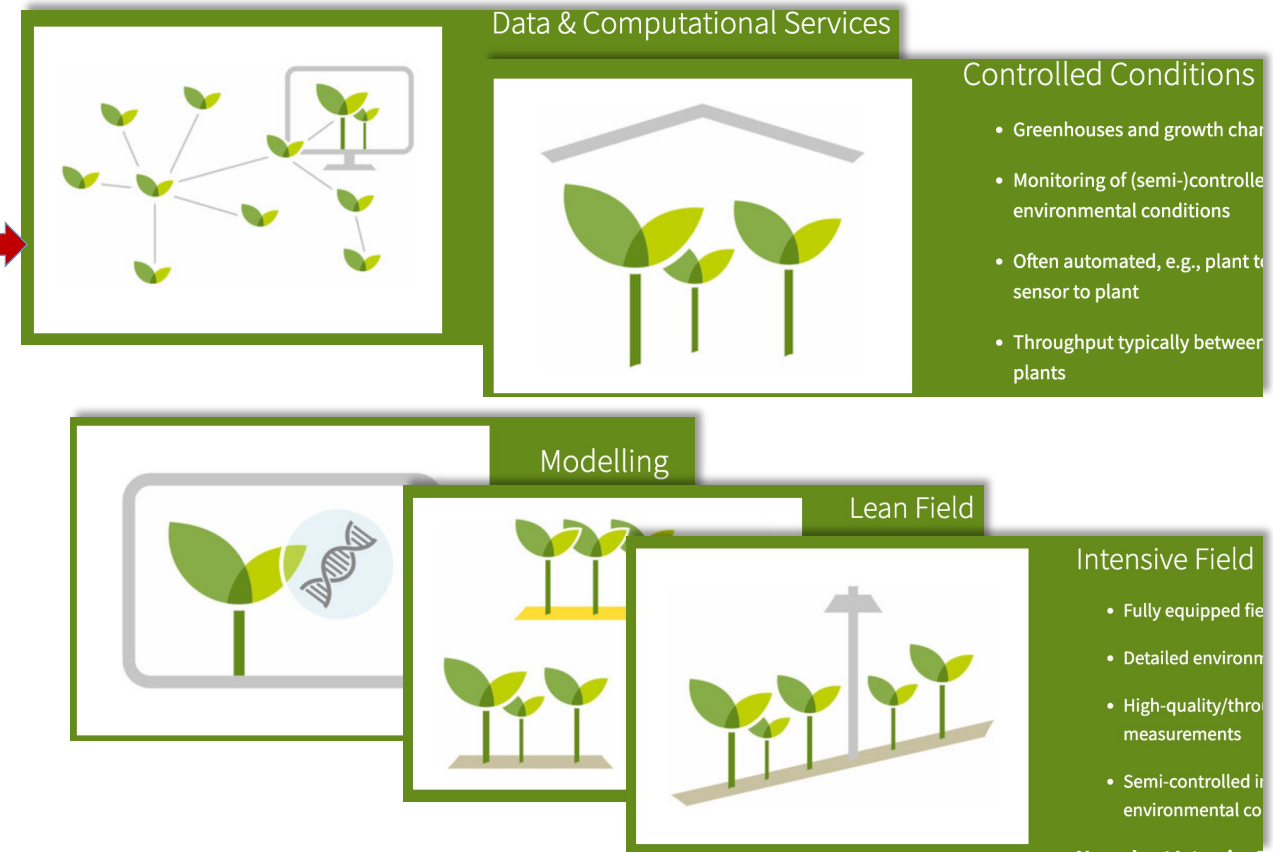
Gathered to solve Pheno to Omic data management and integration

• ELIXIR

- European Infrastructure for life sciences data
- Slovenia, France, Germany, Portugal, UK, Belgium, Italy, NL...
- <https://elixir-europe.org/communities/plant-sciences>
- FAIR data management, software, training, ...

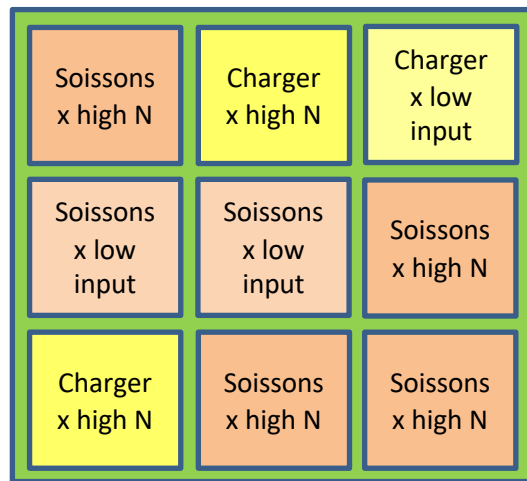
• EMPHASIS IPPN

- European Infrastructure for Plant Phenotyping
- France, Germany, Belgium, UK, ...
- <https://www.plant-phenotyping.eu>



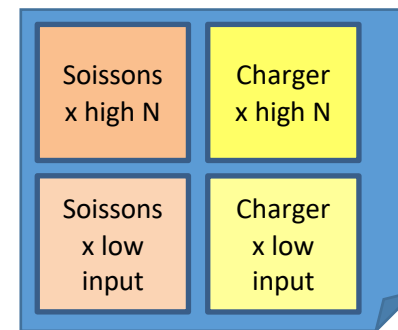


« Raw » data, pheno/env measurement, variables



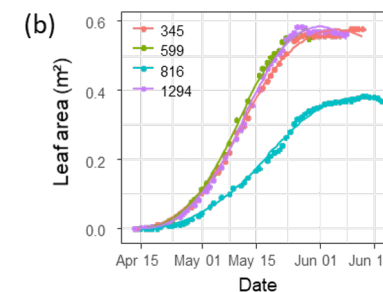
Derivation, Reduction

« Derived » data, indicators



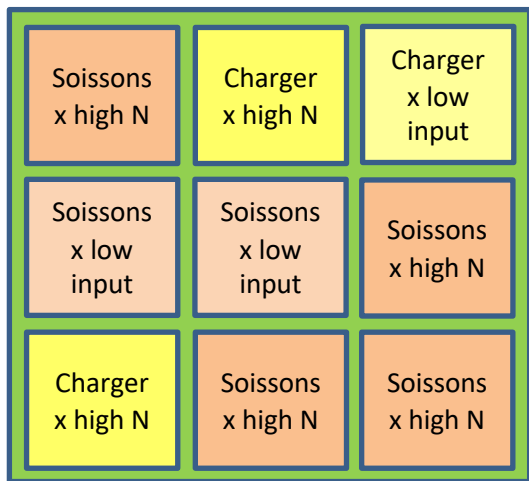
Genotype	Treatment	N input	Date	Rep	Fusariose
Soissons	low input	15,3225129	15/11/2011	1	5
Soissons	low input	15,3430556	16/11/2011	2	7

Genotype	Treatment	Fusariose
Soissons	low input	6



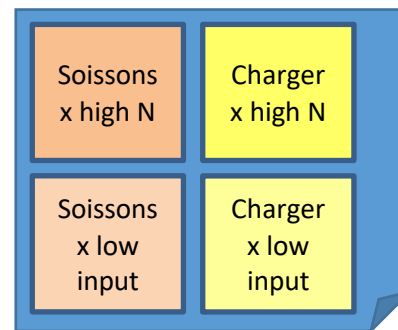


« Raw » data, pheno/env measurement, variables

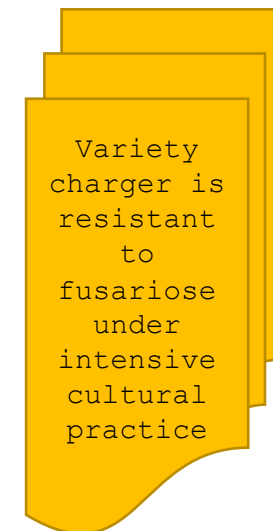


Derivation, Reduction

« Derived » data, indicators

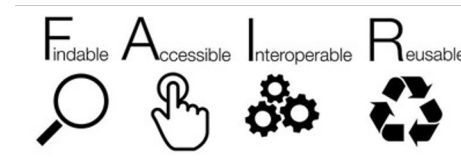
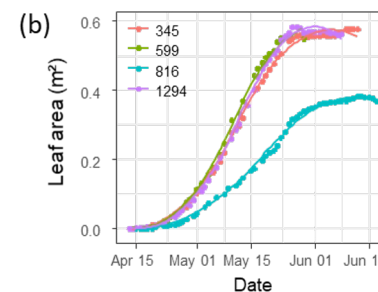


Publication



Genotype	Treatment	N input	Date	Rep	Fusariose
Soissons	low input	15,3225129	15/11/2011	1	5
Soissons	low input	15,3430556	16/11/2011	2	7

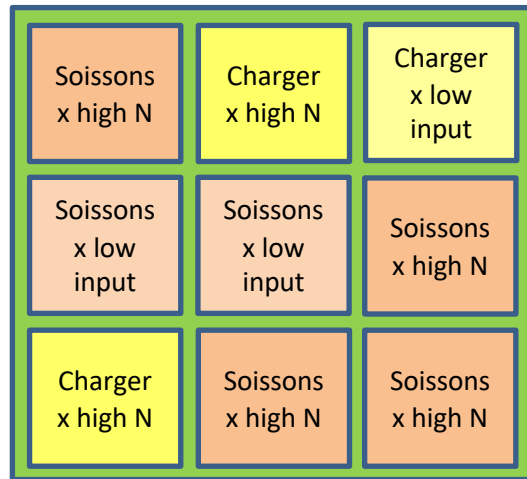
Genotype	Treatment	Fusariose
Soissons	low input	6



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

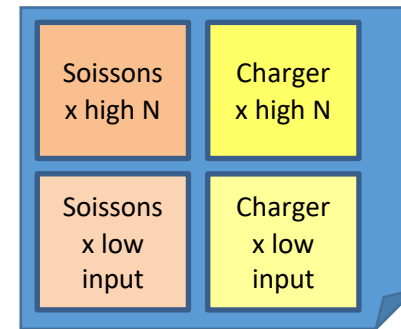


« Raw » data, pheno/env measurement, variables

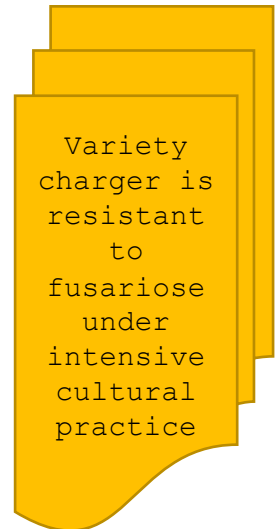


Derivation, Reduction

« Derived » data, indicators



Publication



Genotype	Treatment	N input	Date	Rep	Fusariose
Soissons	low input	15,3225129	15/11/2011	1	5
Soissons	low input	15,3430556	16/11/2011	2	7

Genotype	Treatment	Fusariose
Soissons	low input	6

VARIABLES

- Raw Measures Pheno & Env
- Data cleaning
- Traceability, Reproducibility & Provenance

INDICATORS

- New computation for each scientific question
- One experimental dataset → many derived dataset

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

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

Technical

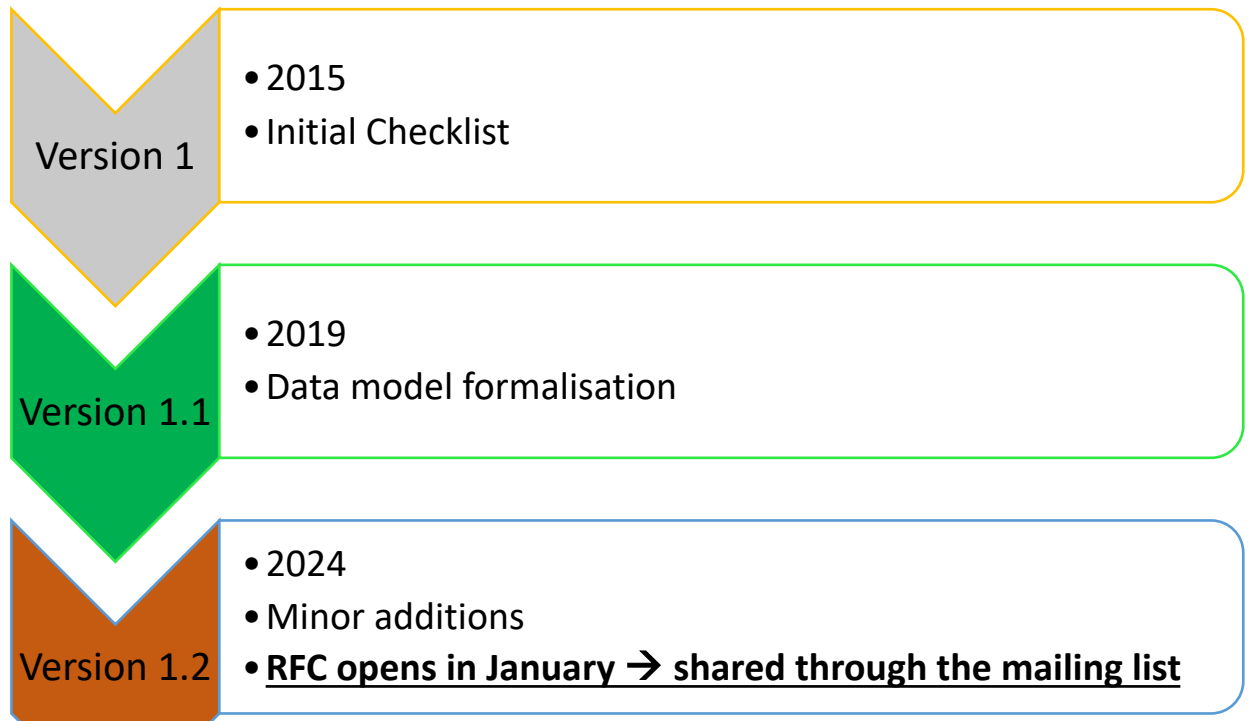
- Data integration and sharing
- Interoperability : tools and systems
 - GA4GH 
 - Breeding API www.brapi.org 
 - Computer scientist driven

Phenotype Structure Standard

Minimal Information About Plant Phenotyping Experiments

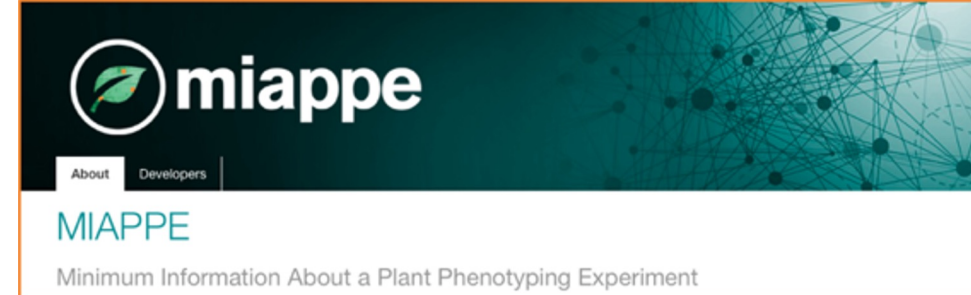


- www.miappe.org
- Crops and woody plants
- Single experiment
- Multilocal multiyear network
- Field
- Greenhouse
- Many stakeholders
 - Elixir, Emphasis, Bioversity CGIAR, North American PPN
- Open Community
 - Request for comments
 - Github Feature requests
 - Mailing lists
 - Meetings & Workgroups



Line #	MIAPPE Check list	Definition	MIAPPE Example	Format	Cardinality
OM-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
OM-2	Investigation unique ID	Identifier comprising the unique name of the institution/database holding the submission of the investigation data, and the accession number of the investigation in that institution.	EM-12345678	Unique identifier	0-1
	Investigation title	Human-readable string summarising the investigation.	Acclimation of Maize to Temperate Climates: Mail-Density Genome-Wide Association Genetics and Density Patterns Revealed Key Genetic Regions, with	Free text (short)	1
Environment					
ENV-1 Non-exhaustive list of Environment Parameters.					
ENV-2	Environment parameters				
Growth facility					
ENV-3	Air temperature	List of hourly air temperature throughout the experiment.	22 °C	Numeric	
ENV-4	Organ temperature	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	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_0007036'	
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'	

• 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
- **Computer scientist friendly**
 - Model, implementations, formalisation
- **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
- Template Excel

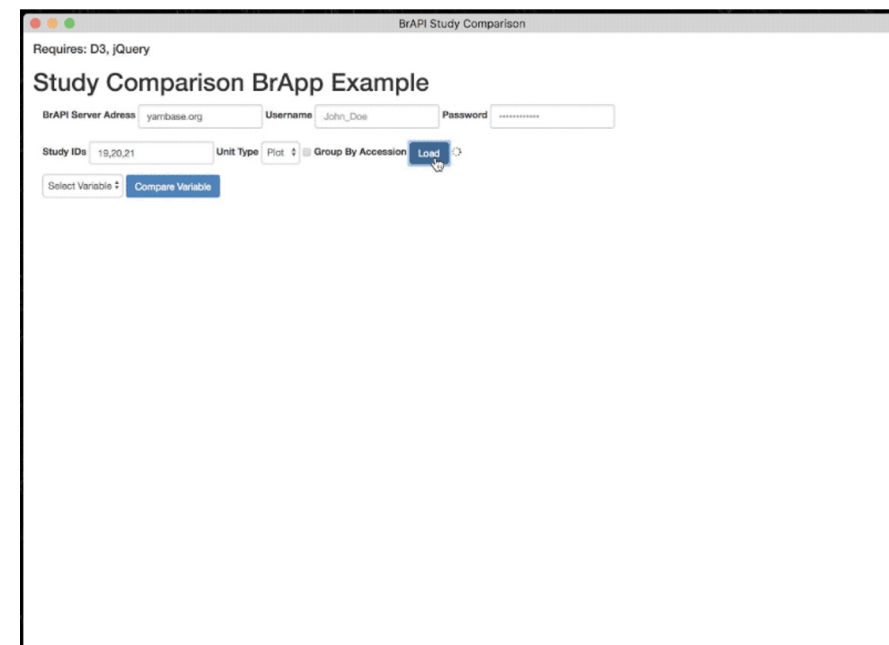
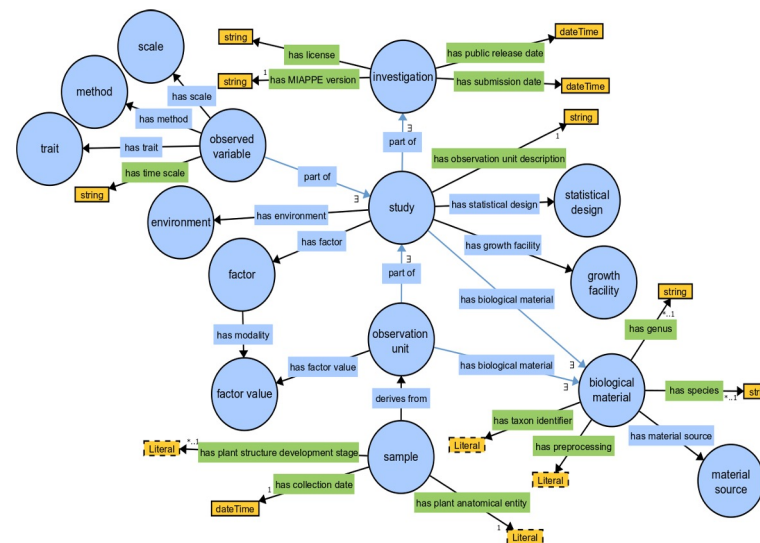
Web Services

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



Data repositories

- Any BrAPI compliant DB (GnpIS, PHIS, PIPPA, ...)
- Generic data repositories (Dataverse, Zenodo, e!Dale, ...)



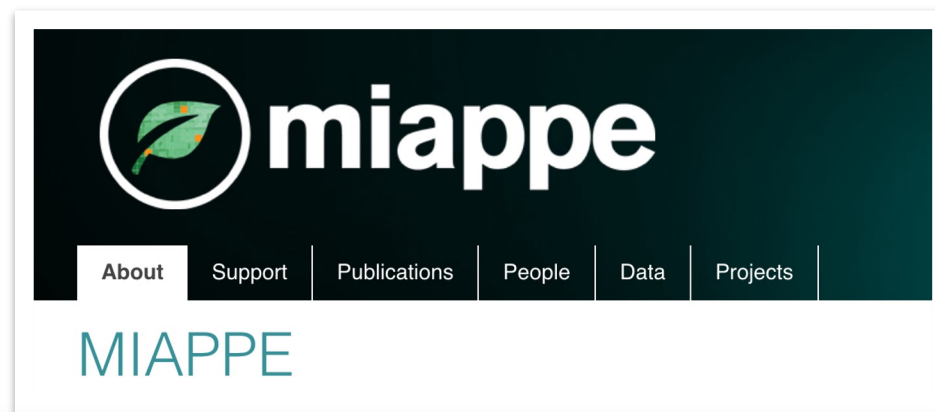
• MIAPPE Specifications

- www.miappe.org

MIAPPE primers

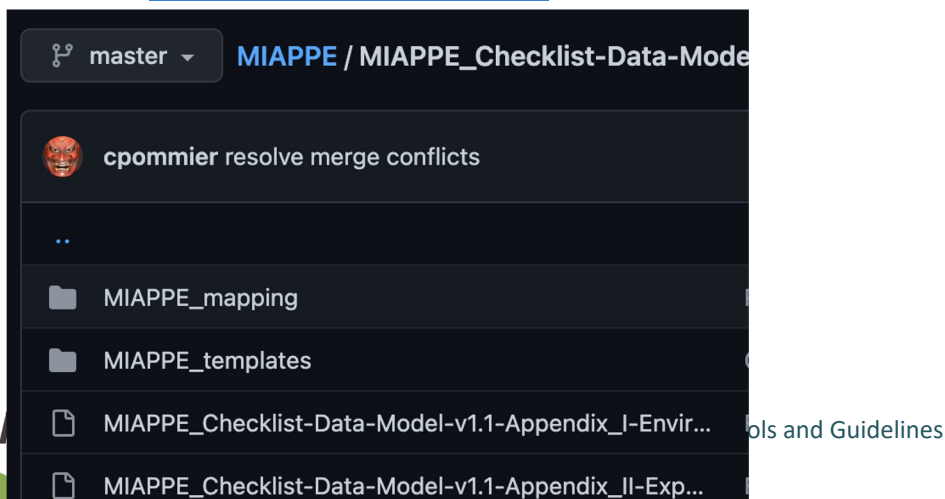
See the [support page](#) for full informations

- The latest specifications, [data model overview](#)
- The latest specifications [field list with description](#)



- Github

- https://github.com/MIAPPE/MIAPPE/tree/master/MIAPPE_Checklist-Data-Model-v1.1



New Phytologist

Methods | [Open Access](#) |

Enabling reusability of plant phenomic datasets with MIAPPE 1.1

Evangelia A. Papoutsoglou✉, Daniel Faria, Daniel Arend, Elizabeth Arnaud, Ioannis N. Athanasiadis, Inês Chaves, Frederik Coppens, Guillaume Cornut, Bruno V. Costa, Hanna Ćwiek-Kupczyńska, Bert Droesbeke, Richard Finkers, Kristina Gruden, Astrid Junker, Graham J. King, Paweł Krajewski, Matthias Lange, Marie-Angélique Laporte, Célia Michotey, Markus Oppermann, Richard Ostler, Hendrik Poorter, Ricardo Ramírez-Gonzalez, Živa Ramšak, Jochen C. Reif, Philippe Rocca-Serra, Susanna-Assunta Sansone, Uwe Scholz, François Tardieu, Cristobal Uauy, Björn Usadel, Richard G. F. Visser, Stephan Weise, Paul J. Kersey, Célia M. Miguel, Anne-Françoise Adam-Blondon, Cyril Pommier✉ ... [See fewer authors](#) ^

First published: 14 March 2020 | <https://doi.org/10.1111/nph.16544> | Citations: 10

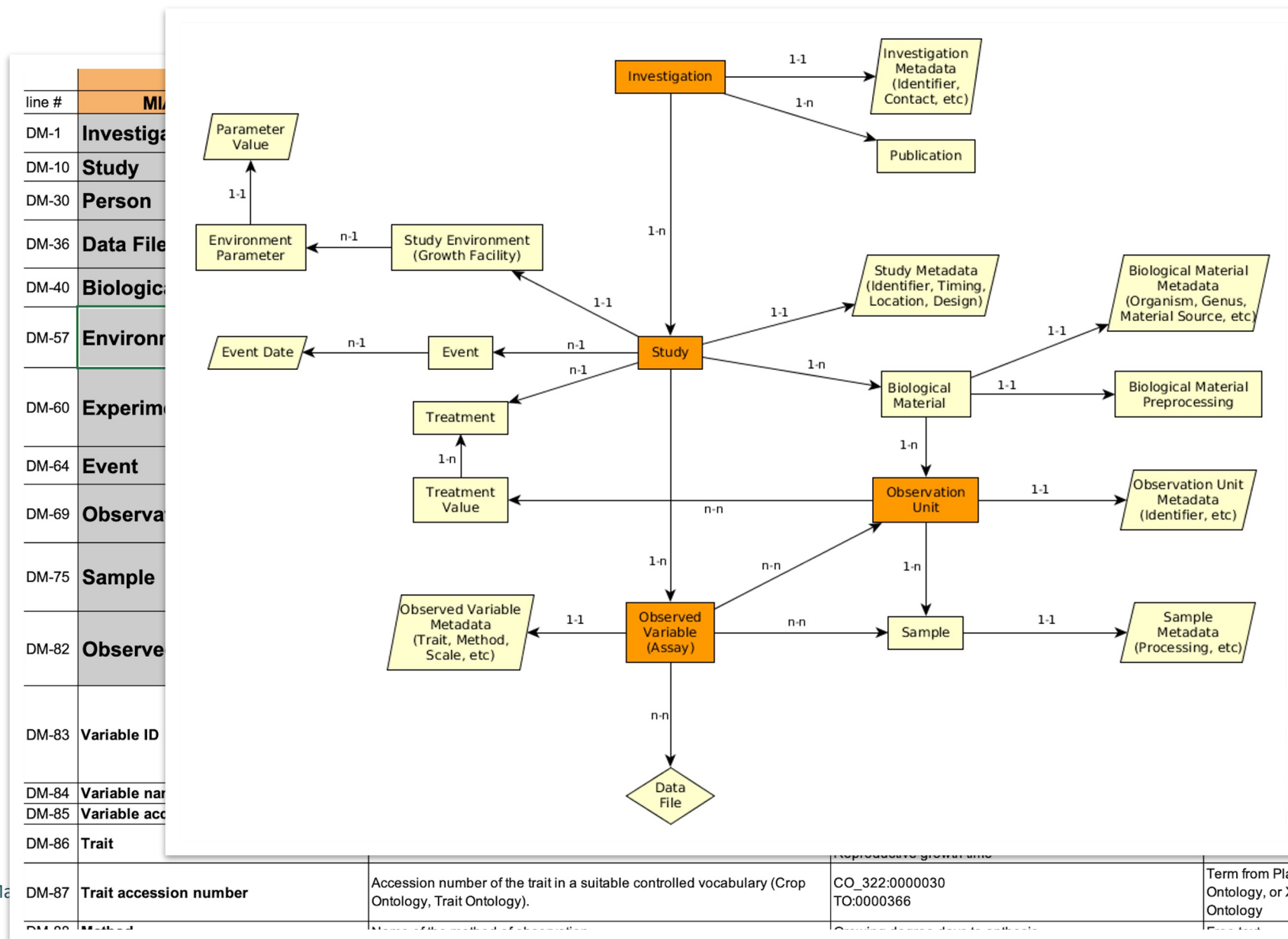
MIAPPE Specifications

- Specification table
- Sections
- Metadata Fields

line #	MIAPPE Check list	MIAPPE			Cardinality
		Definition	Example	Format	
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-10	Study	A study (or experiment) comprises a series of assays (or measurements) of one or more types, undertaken to answer a particular biological question.			1+ per investigation
DM-30	Person	A human involved in the investigation or specifically any of its studies.			1+ per investigation / 0+
DM-36	Data File	A file or digital object holding observation data recorded during one or more assays of the study, typically in tabular form. Multiple data files may be provided per study, and each file can include observations for several observation units and several observed variables.			0+ per study
DM-40	Biological Material	The biological material being studied (e.g. plants grown from a certain bag or seed, or plants grown in a particular field). The original source of that material (e.g., the seeds or the original plant cloned) is called the material source, which, when held by a material repository, should have its stock identified.			1+ per study; 0+ per observation unit
DM-57	Environment	Environmental parameters that were kept constant throughout the study and did not change between observation units or assays. Environment characteristics that vary over time, i.e. environmental variables, should be recorded as Observed Variables (see below).			0-1 per study
DM-60	Experimental Factor	The object of a study is to ascertain the impact of one or more factors on the biological material. Thus, a factor is, by definition a condition that varies between observation units, which may be biotic (pest, disease interaction) or abiotic (treatment and cultural practice) in nature. Depending on the level of the data, an experimental factor can be either "what is the factor applied to the plant" (i.e. Unwatered), or the "environmental characterisation" (i.e. if no rain on unwatered plant : Drought ; if rain on unwatered plant: Irrigated)			0+ per study; 0+ per observation unit
DM-64	Event	An event is discrete occurrence at a particular time in the experiment (which can be natural, such as rain, or unnatural, such as planting, watering, etc). Events may be the realization of Factors or parts of Factors, or may be confounding to Factors. Can be applied at the whole study level or to only a subset of observation units.			0+ per study/observation unit
DM-69	Observation Unit	Observation units are objects that are subject to instances of observation and measurement. An observation unit comprises one or more plants, and/or their environment. There can be pure environment observation units with no plants. Synonym: Experimental unit.			1+ per study
DM-75	Sample	A sample is a portion of plant tissue harvested, non-harvested or extracted from an observation unit for the purpose of sub-plant observations and/or molecular studies. A sample must be used when there is a physical sample that needs to be stored and traced. Otherwise, observations made at the sub-plant level should be recorded as plant level observations using the observed variables to characterize the object of the observation (e.g. Berry sugar content, Fruit weight, Grain Protein content, Leaf 1 width, Leaf 2 width, Leaf 2 length).			0+ per observation unit
DM-82	Observed Variable	An observed variable describes how a measurement has been made. It typically takes the form of a measured characteristic of the observation unit (plant or environmental trait), associated to the method and unit of measurement. Multiple variables with the same combination of trait, method and scale can be used in association with different plant parts (leaf 1, leaf 2), when this distinction is necessary for observations referring to different parts of the same observation unit.			1+ per study
DM-83	Variable ID	Code used to identify the variable in the data file. We recommend using a variable definition from the Crop Ontology where possible. Otherwise, the Crop Ontology naming convention is recommended: <trait abbreviation>_<method abbreviation>_<scale abbreviation>. A variable ID must be unique within a given investigation.	Ant_Cmp_Cday	Unique identifier	1
DM-84	Variable name	Name of the variable.	Anthesis computed in growing degree days	Free text	0-1
DM-85	Variable accession number	Accession number of the variable in the Crop Ontology	CO_322:0000794	Crop Ontology term	0-1
DM-86	Trait	Name of the (plant or environmental) trait under observation	Anthesis time Reproductive growth time	Free text	1
DM-87	Trait accession number	Accession number of the trait in a suitable controlled vocabulary (Crop Ontology, Trait Ontology).	CO_322:0000030 TO:0000366	Term from Plant Trait Ontology, Crop Ontology, or XML Environment Ontology	0-1

MIAPPE Specifications

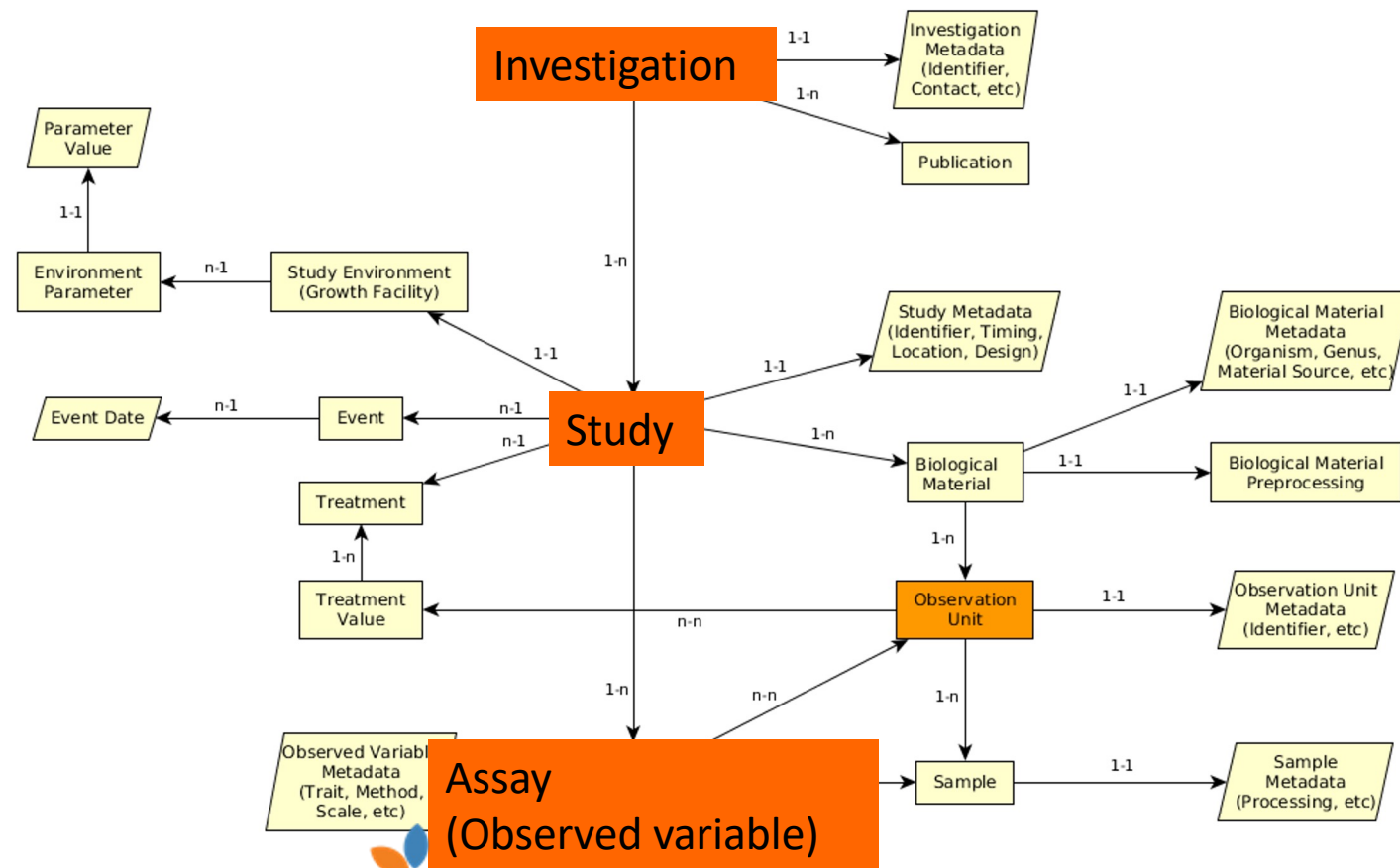
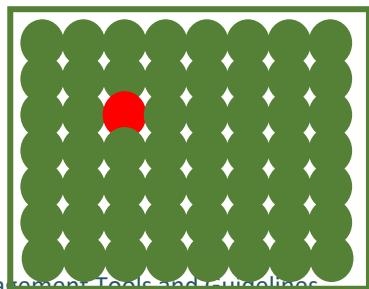
- Specification table
- Sections
- Metadata Fields
- Linked and organized



line #	Metadata Field
DM-1	Investigation
DM-10	Study
DM-30	Person
DM-36	Data File
DM-40	Biological Material
DM-57	Environment
DM-60	Experiment
DM-64	Event
DM-69	Observation Unit
DM-75	Sample
DM-82	Observed Variable
DM-83	Variable ID
DM-84	Variable name
DM-85	Variable accession number
DM-86	Trait
DM-87	Trait accession number

Format	Value
Investigation	1 per
Study	1+ pe
Person	1+ pe
Data File	0+ pe
Biological Material	1+ pe 0+ pe
Environment	0-1 p
Experiment	0+ pe 0+ pe
Event	0+ pe
Observation Unit	1+ pe
Sample	0+ pe
Observed Variable	1+ pe
Variable ID	1
Variable name	0-1
Variable accession number	0-1
Trait	1
Trait accession number	0-1

- **Investigation:** whole dataset
- **Study :** one experiment in one location for one to several year
- **Assay:**
 - trait or indice (Pheno or Env) observed
 - Level + Trait + Method + Scale/Unit
- **Level:**
 - Plant
 - Microplot
 - Block
 - Trial
 - ...



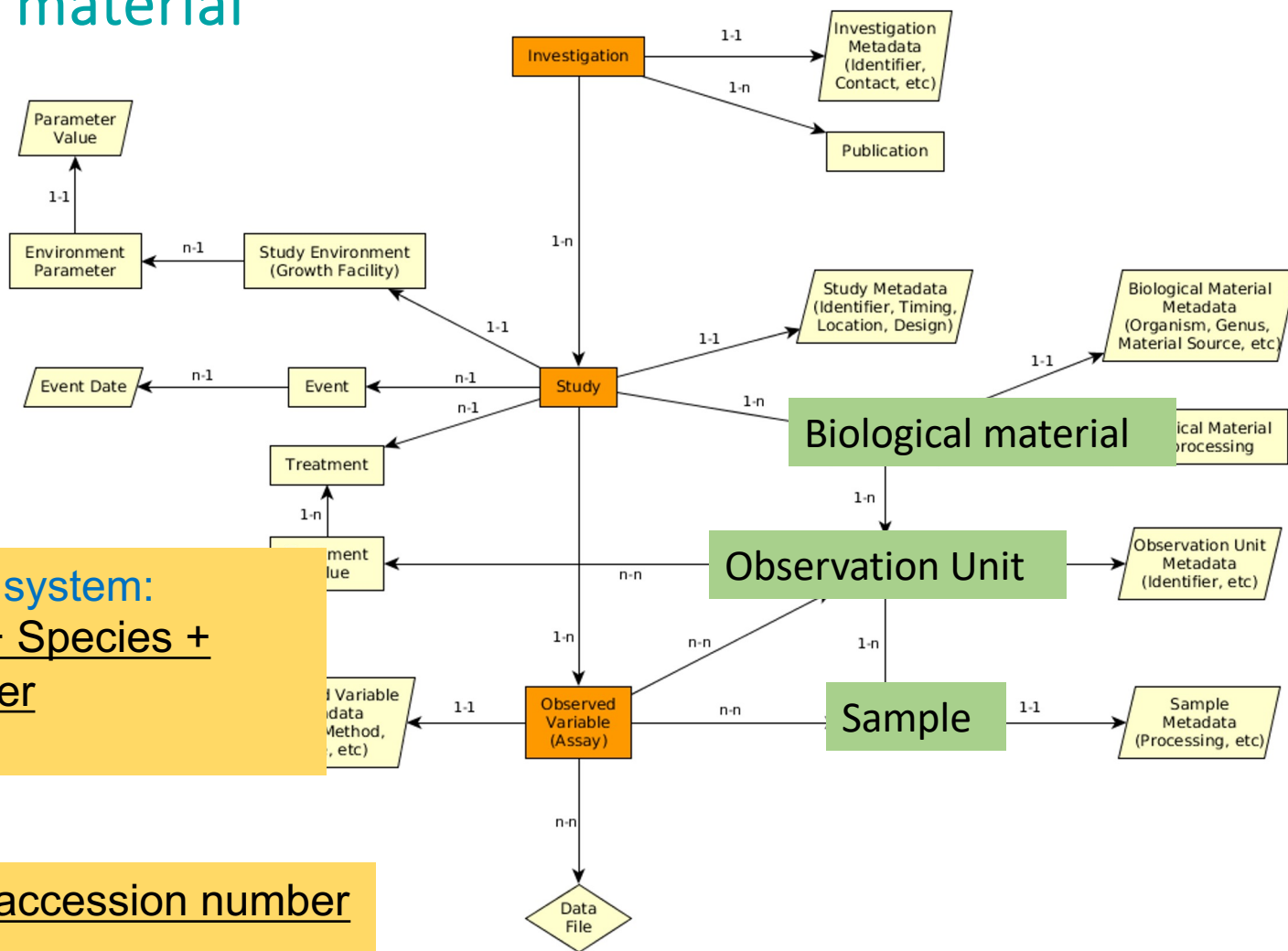
Crop Ontology
for agricultural data

MIAPPE main sections – Biological material

- Plant Material

- Identification
- Description

- Multi Crop Passport Descriptor



MCPD identification system:

- Genebank/Lab + Species + accession number
- DOI

- Lab + internal accession number
- URI

- Lab + internal accession number (mandatory)
- BioSample ID

Material Source: accession, cultivar/variety, region of provenance, laboratory cross, ...

Biological material used in the study: seed lot, cuttings...

Plant Samples used in the study: detached leaves, ...

MIAPPE main sections – Observed Variable (Assay)



- Plant Phenomic Specific Ontology Model:

Phenotyping/environment variable = *Trait + Method + Unit/Scale*

Trait

+

Method

+

Unit

M1: Total height

M2: First tassel branch

M3: Last expanded leaf

M4: Youngest growing leaf

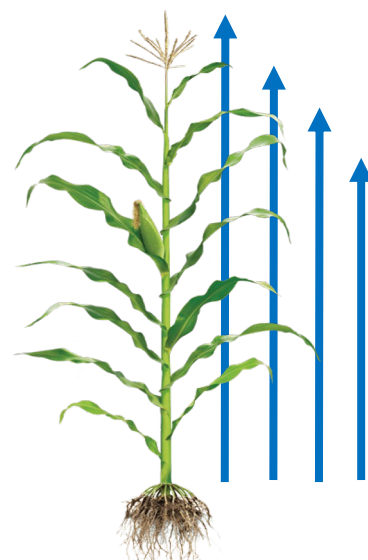
U1: cm

U2: mm

T1: Plant Height

M5: Highest pixel corresponding to plant

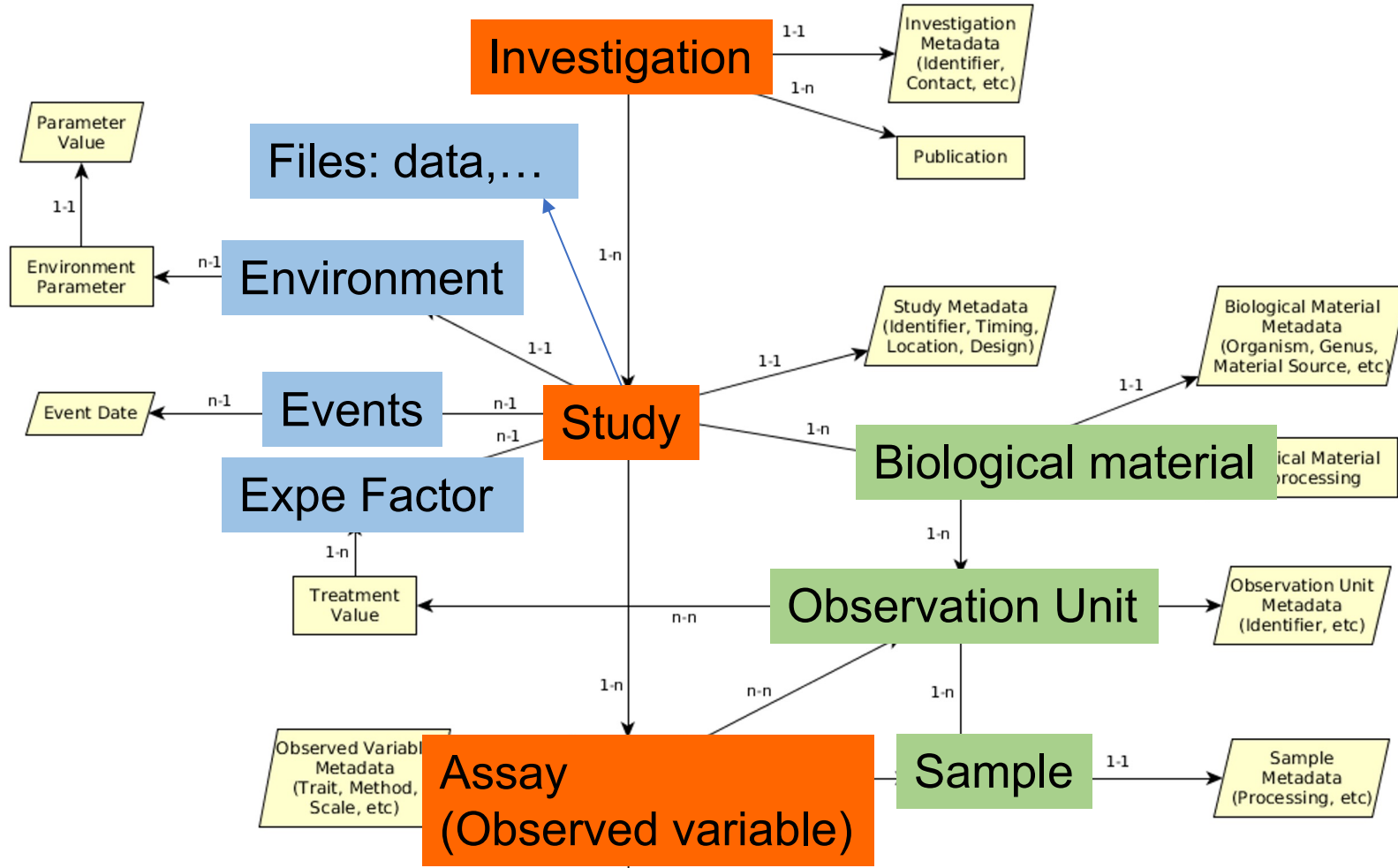
U3: pixel



guidelines

Slide:
L Cabrera-Bosquet

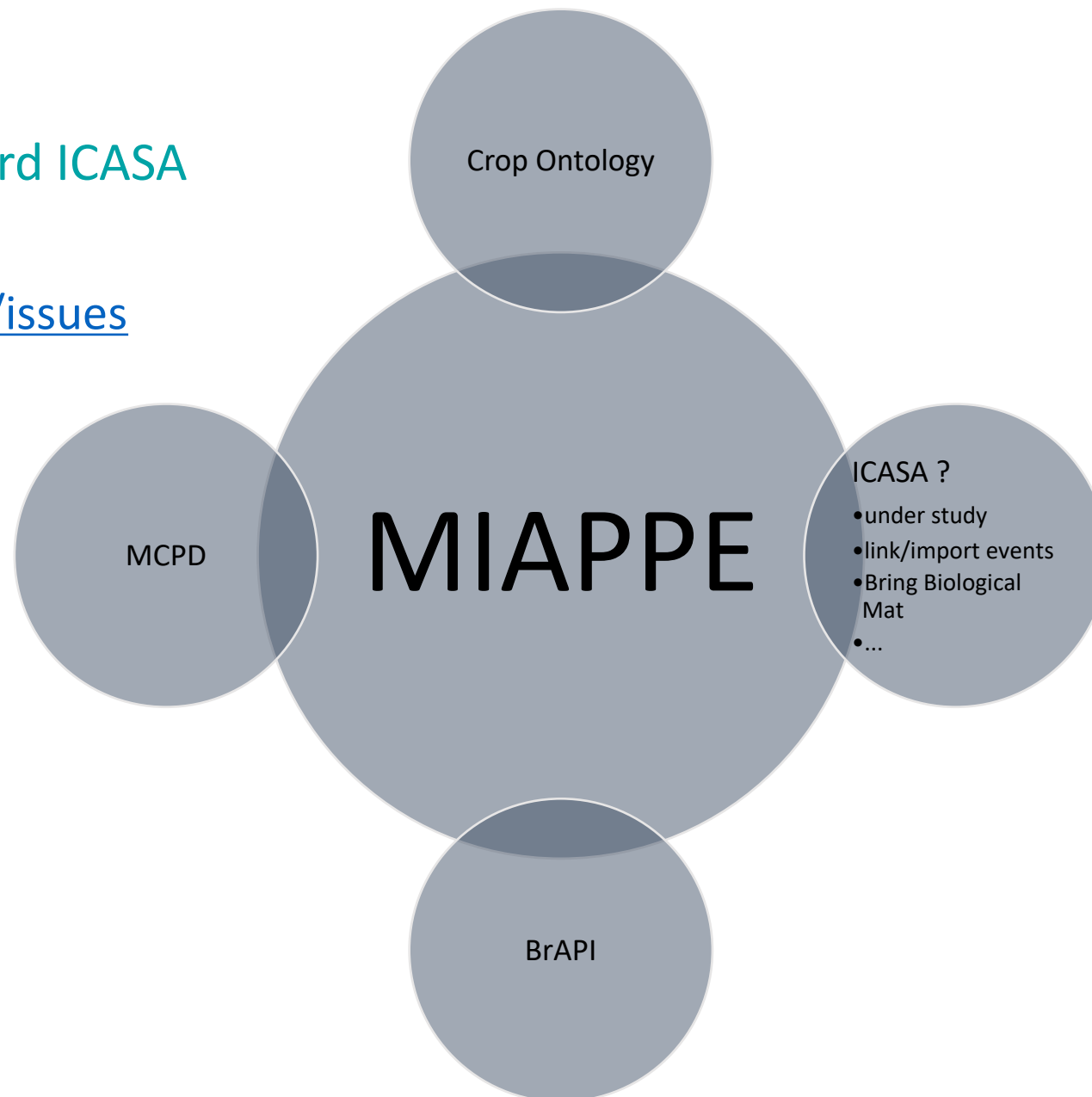
- MIAPPE – Other Important sections



- MIAPPE Overview
Data file content
- Any format (Near Infra Red Spectrum, Images, Image Archives references,)
- Mostly tabular

A	B	C	D	E	F	G	H	I
Accession Number	Trial Site	Campaign	Circum1: Tree circumference at 1 year	Date [Circum1]	Height1: Tree total height at 1 year	Date [Height1]	Shoots3: Number of resprouts at 3 years	Date [Shoots3]
661300270	Ardon	2004	45.645632645603683	12/01/2004	284.3	12/01/2004		
661300270	Ardon	2005					14.630625	12/05/2005
661300444	Ardon	2004	38.96112577281653	12/01/2004	228.8	12/01/2004		
661300444	Ardon	2005					8.5030559999999991	12/05/2005
661300312	Cavallermaggiore	2004	52.4	01/01/2004	249.9	01/01/2004		
661300312	Cavallermaggiore	2005					12.981609000000001	01/05/2005
661300371	Cavallermaggiore	2004	45.74	01/01/2004	230.2	01/01/2004		
661300371	Cavallermaggiore	2005					10.3041	01/05/2005
661300487	Cavallermaggiore	2004	72.52	01/01/2004	309.8	01/01/2004		
661300487	Cavallermaggiore	2005					10.679823999999998	01/05/2005
661300585	Cavallermaggiore	2004	71.739999999999995	01/01/2004	305.7	01/01/2004		
661300585	Cavallermaggiore	2005					10.956100000000001	01/05/2005
661300468	Headley	2004	45.27	01/01/2004		247	01/01/2004	
661300468	Headley	2005					15.888196000000002	01/05/2005
661300469	Headley	2004	70.930000000000007	01/01/2004		313	01/01/2004	
661300469	Headley	2005					13.271448999999999	01/05/2005
661300533	Headley	2004	57.67	01/01/2004	258.8	01/01/2004		

- Recommended data file format
- Extend standard interoperability toward ICASA
- Github issues
 - <https://github.com/MIAPPE/MIAPPE/issues>

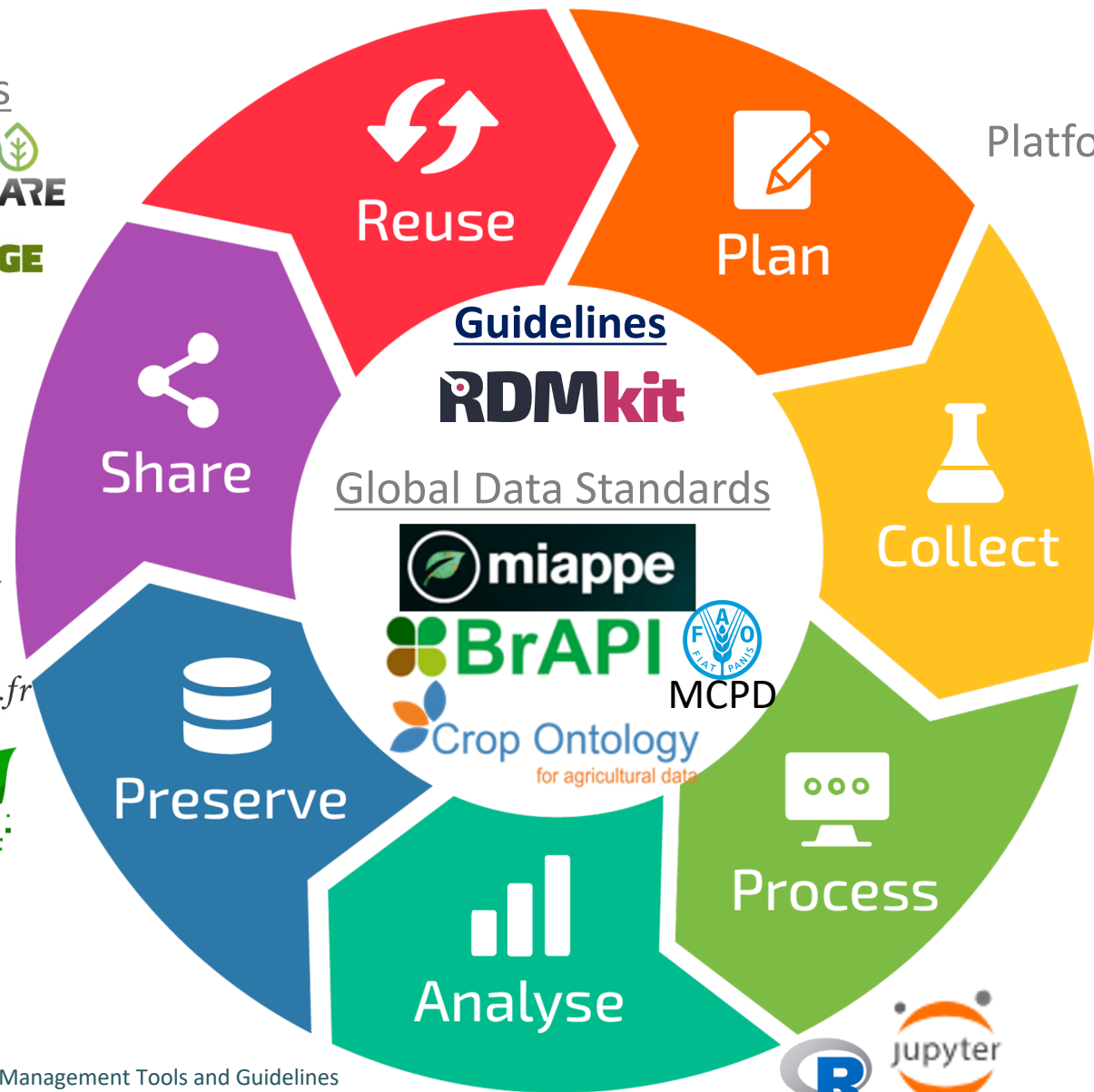
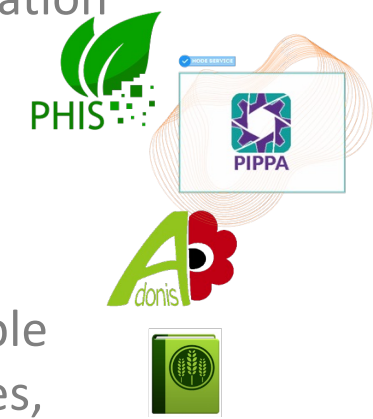




Data portals

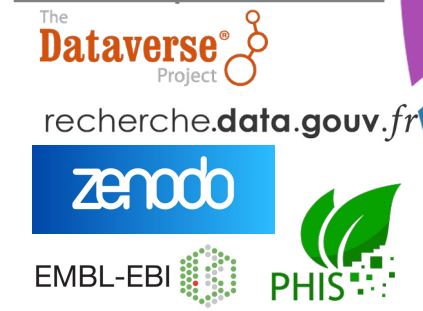


Platform Information Systems



Portable devices, sensors, ...

Data repositories

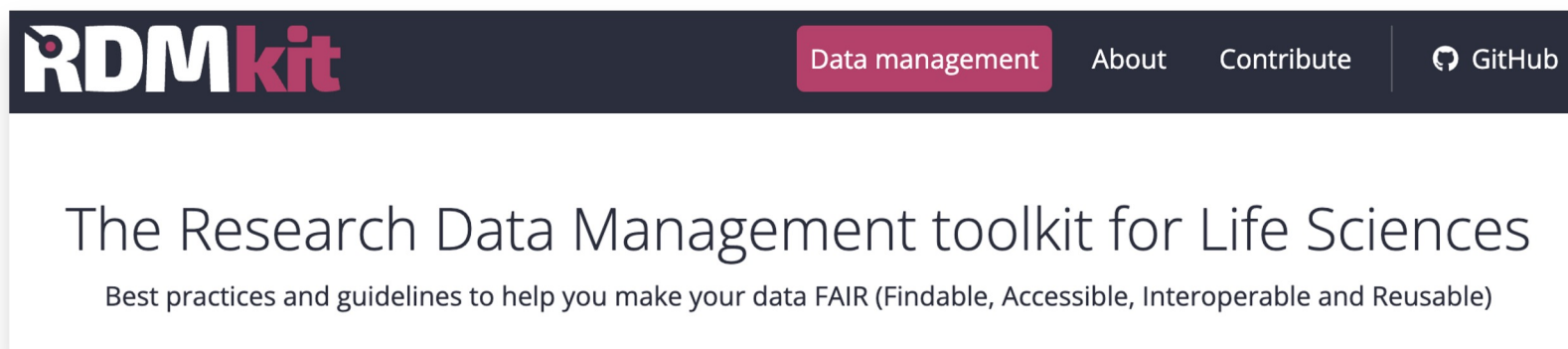


Scripts and Workflows



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 in numbers



- Contributors are experts in RDM and/or in scientific domain in ELIXIR and beyond
- Managed by an editorial board
 - ELIXIR members from several Nodes
 - 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



Considerations

- Did you collect the metadata for the identification of your plant material according to the recommendation provided in the [above section](#)?
- Have you documented your phenotyping and environment assays (i.e. measurement or computation methodology based on the trait, method, scale triplet) both for direct measures (data collection) and computed data (after data processing or analysis)?
 - Is there an existing [Crop Ontology](#) for the species you experiment and does it describe your assay? If not, have you described your data following the trait, method, scale triplet?
- Do you have your own system to collect data and is it compliant with the [MIAPPE](#) standard?
- Are you exchanging data with individual researchers?
 - In what media is data being collected?
 - Is the data described in a [MIAPPE](#) -compliant manner?
- Are you exchanging data across different data management platforms?
 - Do these platforms implement the Breeding API [BrAPI](#) specification?
 - If not, are they MIAPPE-compliant and do they enable automated data exchange?

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

Solutions

Checklists and ontologies

- The metadata standard applicable to plant phenotyping experiments is [MIAPPE](#).
 - There is a section dedicated to the identification of plant biological materials that follows [Multi-Crop Passport Descriptor \(MCPD\)](#) described [above](#).
 - There is a section to describe the phenotyping assays based on the [Crop Ontology](#) recommendations.

Data management practices in a domain

- what aspects should be taken into account (considerations)
- available solutions
- links to tools and resources explained in clear context (when to use them and for what purpose)

Plant Pages

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

RDMkit Data management About Contribute GitHub Search RDMkit

Data management

- Data life cycle
- Your role
- Your domain
- Your tasks
- Tool assembly**
- COVID-19 Data Portal
- CSC
- Galaxy
- IFB
- Marine Metagenomics
- MOLGENIS
- NeLS
- OMERO
- Plant Genomics
- Plant Phenomics
- TransMed

Tool assembly

Tool Assemblies are examples of combining tools to cover data management tasks across several stages of the data life cycle. These can be tools that one or several communities combine to support RDM that can be picked up or accessed and used by others. The assemblies are aimed for users in a specific location and/or for users within a specific domain.

Filter by affiliation Choose... Search Type here...

COVID-19 Data Portal

The COVID-19 Data Portal brings together relevant datasets for sharing and analysis to accelerate coronavirus research.

Related pages

- Your tasks
- Data sensitivity
- Existing data
- Data publication
- Data analysis
- Your domain
- Human data

Affiliations: elixir

CSC

The Center of Science (CSC) provides high-quality ICT expert services for researchers in Finland and their collaborators.

Related pages

Affiliations: + csc elixir

IFB

The French Bioinformatics Institute (IFB) offers IT infrastructure and bioinformatics expertise to support researchers in Life Sciences.

Related pages

Affiliations: elixir

Tool assemblies for plant data

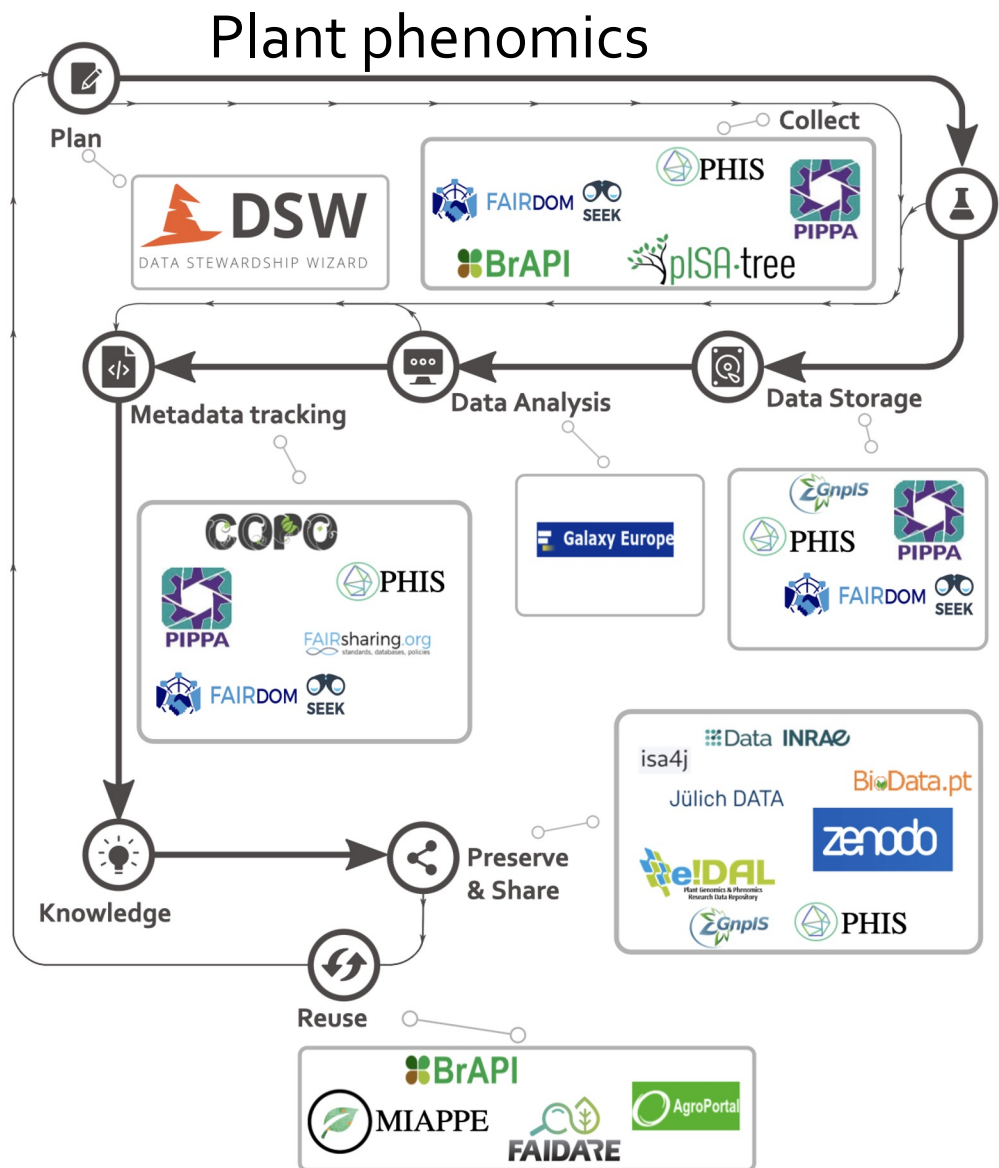


Figure 1. The plant phenomics tool assembly.

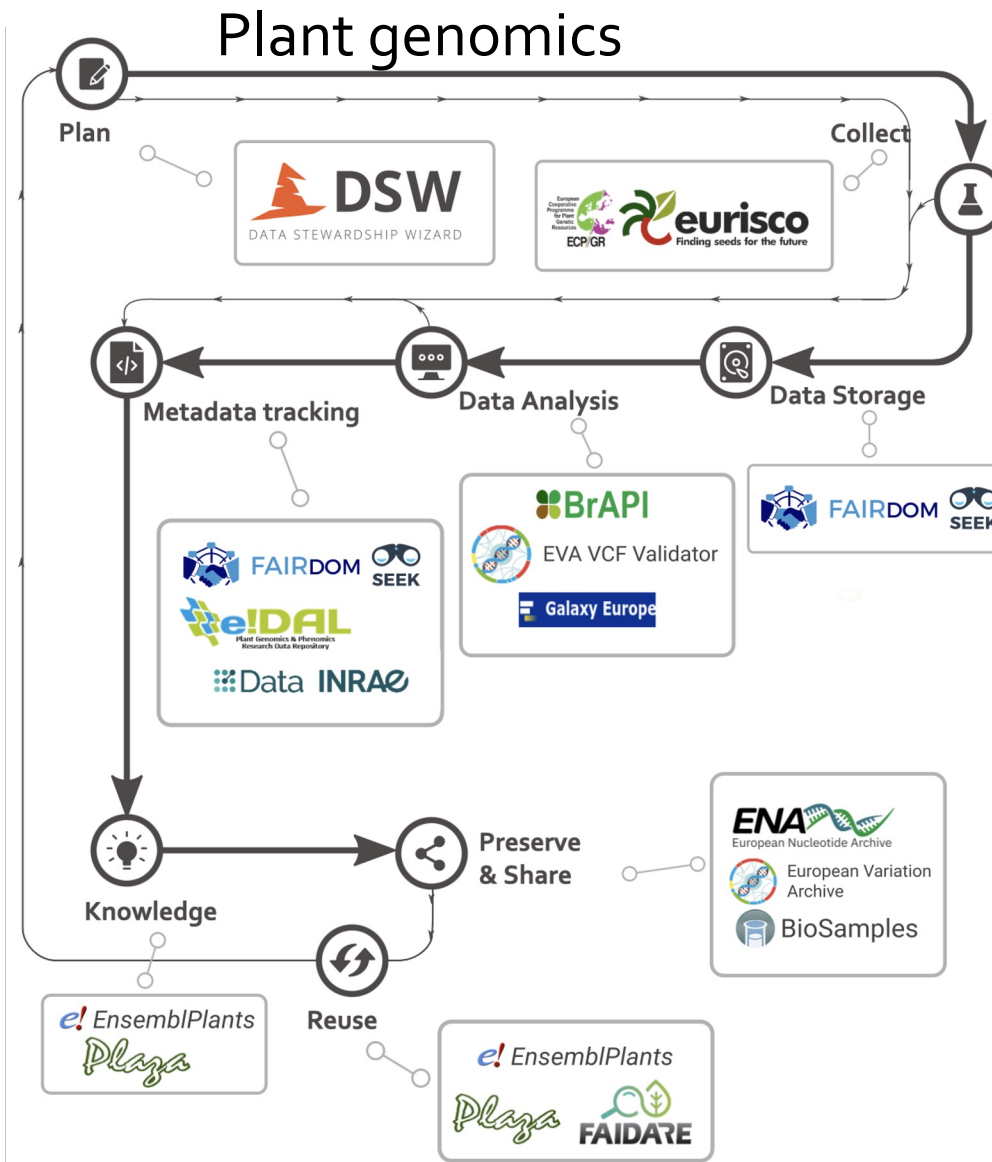


Figure 1. The plant genomics tool assembly.

Tool assemblies for plant data

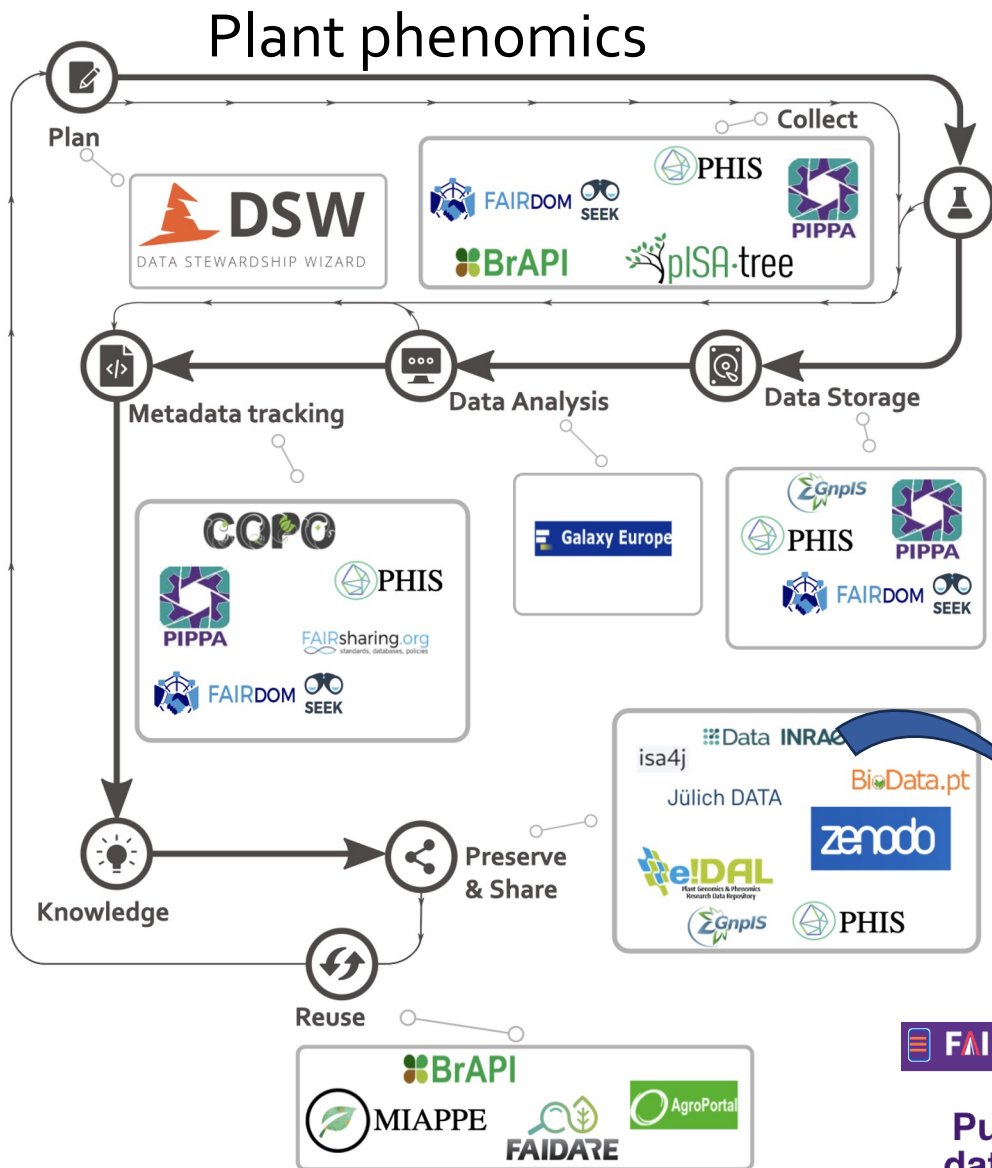


Figure 1. The plant phenomics tool assembly.

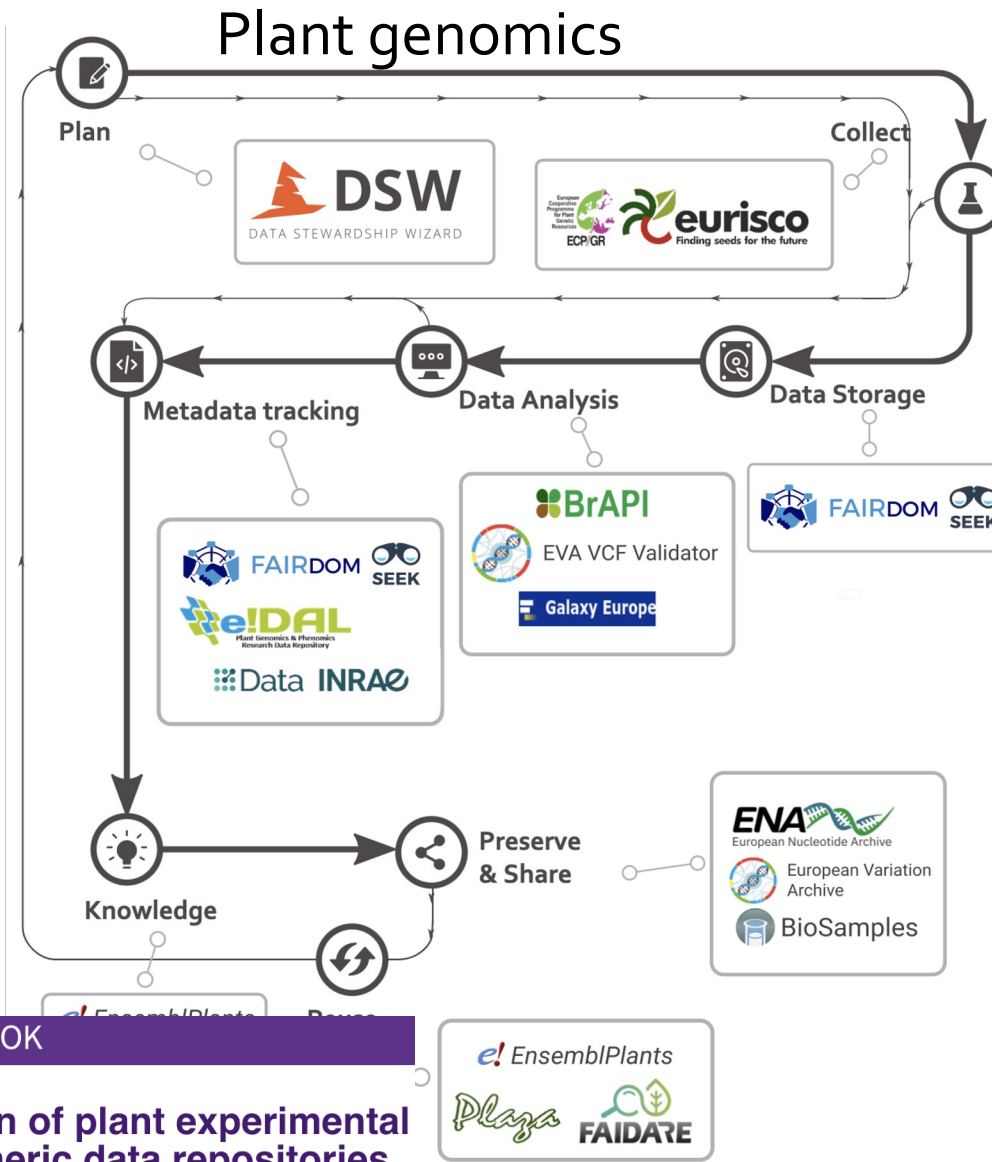


Figure 1. The plant genomics tool assembly.

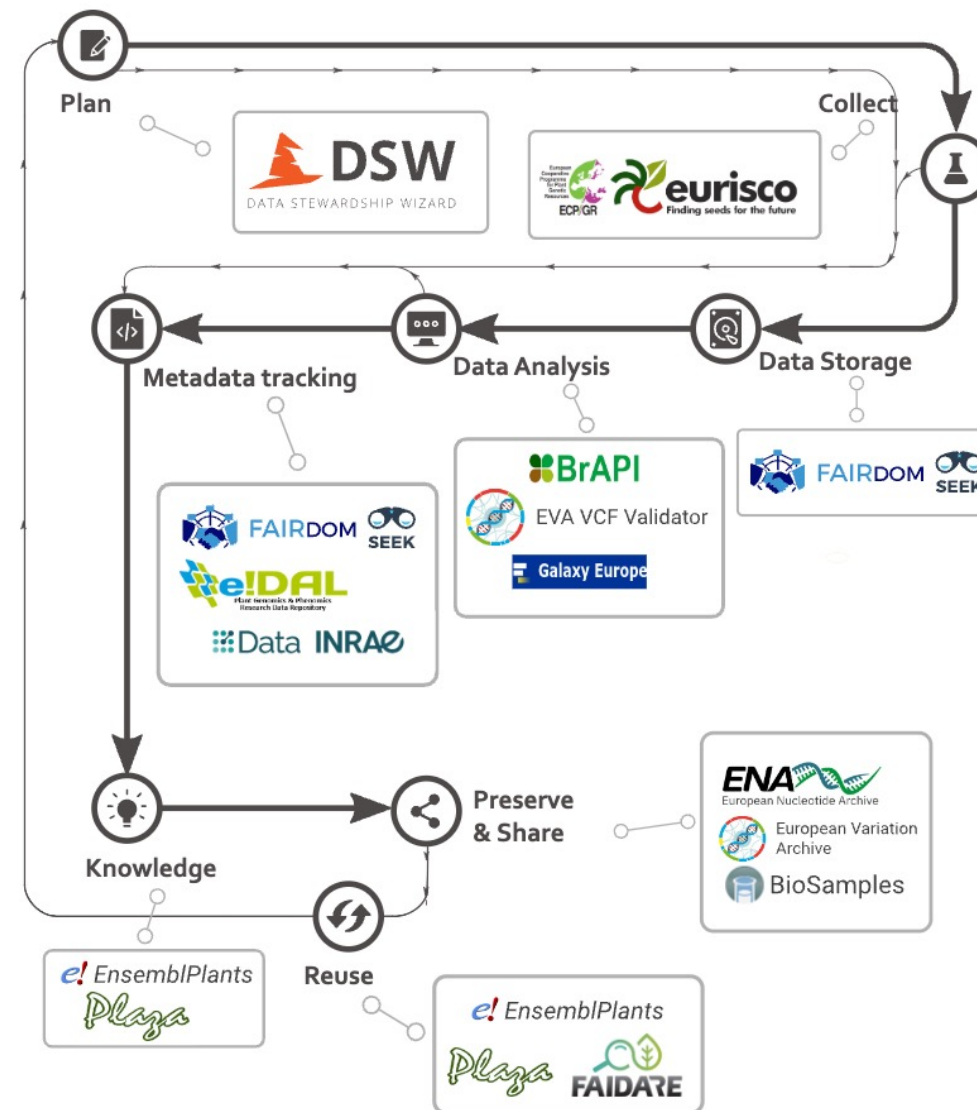
FAIRCOOKBOOK

Publication of plant experimental data in generic data repositories

Publish Sequence and Variation @ EMBL-EBI

- Step by step procedure
- Publish data @ EBI
- Ensure interoperability from Pheno to Geno

https://rdmkit.elixir-europe.org/plant_genomics_assembly



Publish in generic data repositories

- **Dataverse**

- Zenodo
- e!Dale

- **PAG Computer Demo**

- Saturday the 13th at 13:30 Town and Country D

Details

Location: Town and Country D

Date: Saturday, Jan 13 1:30 PM

Duration: 2 hours 10 minutes

Presentation



1:30 PM Long Term Plant Phenomic Data Sharing in Generic Data Repositories (Zenodo, Dataverse) Using MIAPE

https://rdmkit.elixir-europe.org/plant_phenomics_assembly

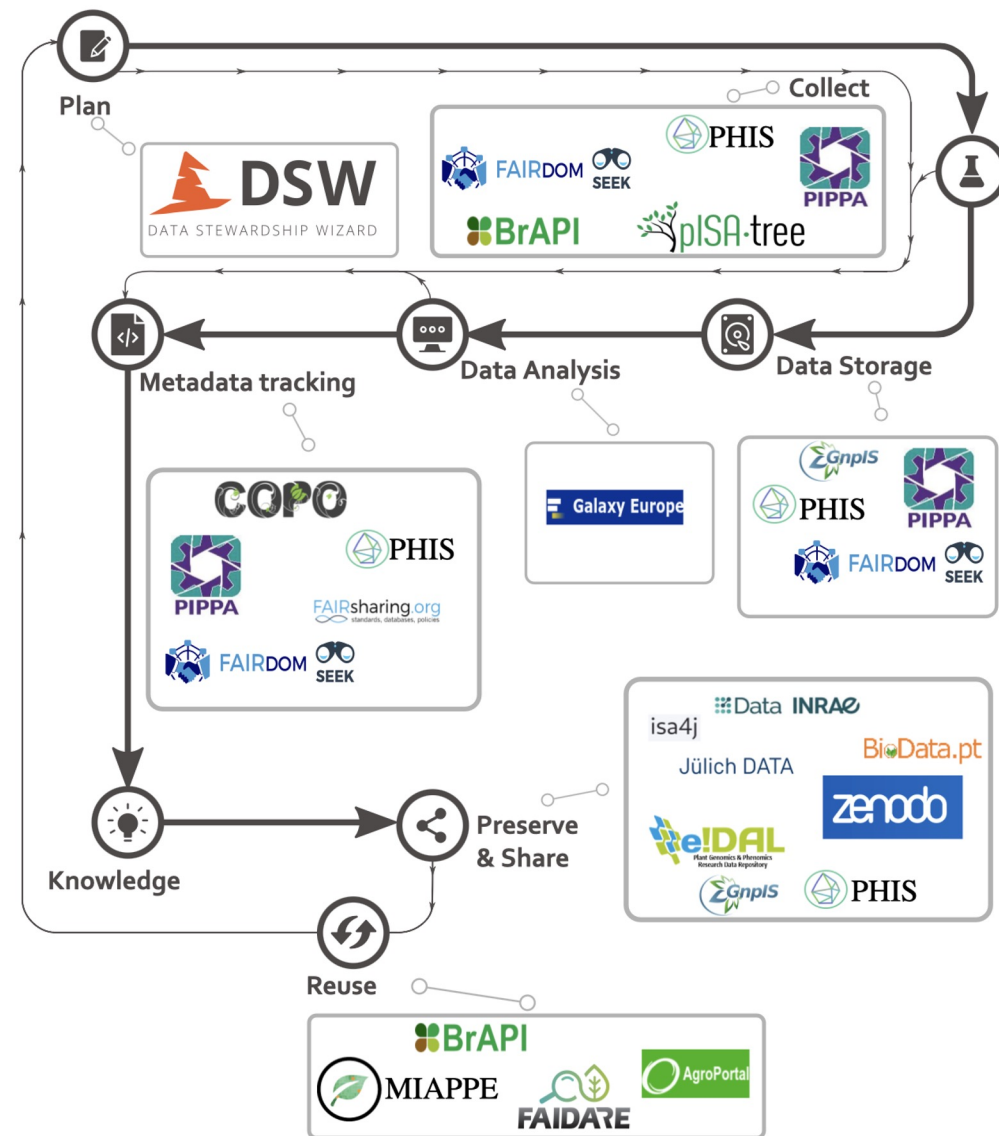


Figure 1. The plant phenomics tool assembly.



Dataverse

<https://faircookbook.elixir-europe.org/content/recipes/reusability/plant-pheno-data-publication.html#step-by-step-process-for-data-submission-and-publication-in-dataverses>

<https://demo.dataverse.org/>

Step 1: Dataset creation: Minimal metadata

Step 1: Dataset creation

Find an appropriate dataverse

You need first to select the appropriate dataverse and/or sub-dataverse for your use case depending on the constraints of your consortium or institute. The guidelines explained below are applicable to all of those dataverse instance. You will find below a non -exhaustive list of dataverse instances. Also, find the right sub-dataverse for your submission, like the one of the research group or the project you belong to: dataverses can contain dataverses.

- recherche.data.gouv.fr (FR)
 - Open to submission from any consortium involving at least one member of
 - Examples: [Data INRAE](#) and [CIRAD](#).
- [Jülich DATA](#) (NRW - DE)
 - Open to submission from any research activity done from partners at Forsch
 - Meant for data and software submissions.
 - Maintained by the central library of FZJ.
 - Examples under subject "[Agricultural Sciences](#)".
- dmpportal.biodata.pt (PT)
 - Open to submission of biological data from Portuguese research & innovati
 - Example: [Plant BioDataVerse](#).

Recherche Data Gouv, l'écosystème au service du partage et de l'ouverture des données de recherche célèbre ses 1 an

Recherche Data Gouv Génération datapaper
(Recherche Data gouv)

Metrics 675,807 Downloads Contact Share

L'entrepôt pluridisciplinaire *Recherche Data Gouv* est une solution souveraine pour le partage et l'ouverture des données de recherche produites par les communautés qui ne disposent pas d'un entrepôt disciplinaire reconnu. Il est basé sur le logiciel libre *Dataverse*. Le dépôt des données doit se faire dans l'espace institutionnel attribué à l'établissement dont relève un des contributeurs. Un espace générique est dédié aux données produites par les établissements ne disposant pas d'espace dédié. Les tests sont à effectuer dans le [bac à sable](#). Retrouver les actualités et événements de la plateforme *Recherche Data Gouv*.

The *Recherche Data Gouv* multidisciplinary repository is a sovereign solution for sharing and opening up data produced by communities that do not have a recognized disciplinary repository. It is based on the *Dataverse* software. Data should be deposited in the space assigned to an institution which one of the contributor belongs to. A generic space is available for data produced by institutions which do not such a space yet themselves. Tests should be performed in the [sandbox](#). The latest news and events on the *Recherche Data Gouv* platform.

sub-dataverse

Data repositories minimal description limitations

- (Too) broad metadata
 - Year, Data type, Author
 - Organism: no controlled vocabulary
 - Keyword, Subject
 - Lacking plant phenomic specific metadata
 - Biological material, from species to genetic resource accession
 - Traits
 - Experiment locations
 - ...
 - Need of dedicated metadata scheme, added to generic data repositories
 - As additional metadata → Filling long list in the repository form is not an option
 - As companion files
- MIAPPE data standard



MIAPPE @ Dataverse

Step 2: Add mandatory metadata for plant phenotyping data

Biological material

Use [BiologicalMaterial.xlsx](#). This spreadsheet contains the following fields: Mandatory fields:

- “Biological material ID” (ex: INRA:W95115_inra_2001): Lot number or material identifier in the data files
- “Material source ID” (ex: INRA:B73_usda) OR “Accession_number” (B73_usda) + “Holding_institute” (ex: INRA)
- Accession Number
- Genus
- Species
- Optional fields:
- “Material source DOI”: accession DOI
- Organism: NCBITAXON:4577
- “Infraspecific name”: variety names, cultivar names, etc...
- Genealogy:
 - Parent1or2_AccessionNumber
 - Parent1or2_TaxonGroup
 - Parent1or2_HoldingInstitutionName
 - Parent1or2_Type (father/mother/undefined)
- All MIAPPE Biological Material fields ([DM-40](#) to [DM-56](#))
- Free input: synonyms, project IDs, any relevant information on the plant material.

Observed variables

Use [ObservedVariables.xlsx](#). This file is needed for the description of phenotyping experiments traits and methods.

Studies or experiments

It is recommended to list the experimentation done in this dataset, including in particular the GPS location, the site name and the environmental parameters which characterize the experimental sites. Use

[Studies.xlsx](#)

How to add metadata files in a Dataverse

- Click on “+ Upload Files” in the “Files” tab
- Add the file(s). Please note that :
 - the file size is limited to 15.0 GB
 - compressed files are automatically decompressed at the time of import
 - tabbed files must use the “,” separator and “UTF-8” encoding to avoid problems during import (see the dedicated section in the [user guide](#))
- Fill in the “Description” field for each added file
- Update the file labels by selecting “File options” > “Tags” for each file

Update the file labels

- Add a custom label, “Biological_Material” or “Observed_Variable” depending on the file type. If the label exists, it will be available in the “File labels” section, otherwise you will have to create it in the “Customize file label” section and apply it.

Add a custom label

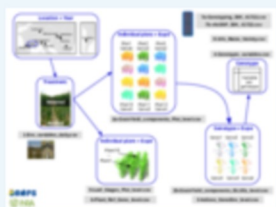
- Save your modifications

Maize Example

DROPS project dataset: <https://doi.org/10.15454/IASSTN>

A multi-site experiment in a network of European fields for assessing the maize yield response to environmental scenarios

Version 4.0



Millet, Emilie J.; Pommier, Cyril; Buy, Mélanie; Nagel, Axel; Kruijjer, Willem; Welz-Bolduan, Therese; Lopez, Jeremy; Richard, Cécile; Racz, Ferenc; Tanzi, Franco; Spitkot, Tamas; Canè, Maria-Angela; Negro, Sandra S.; Coupel-Ledru, Aude; Nicolas, Stéphane D.; Palaffre, Carine; Bauland, Cyril; Praud, Sébastien; Ranc, Nicolas; Presterl, Thomas; Bedo, Zoltan; Tuberosa, Roberto; Usadel, Björn; Charcosset, Alain; van Eeuwijk, Fred A.; Draye, Xavier; Tardieu, François; Welcker, Claude, 2019, "A multi-site experiment in a network of European fields for assessing the maize yield response to environmental scenarios", <https://doi.org/10.15454/IASSTN>, Recherche Data Gouv, V4, UNF:6:zS2/ccOQxFrKIUt+1S0Cvg== [fileUNF]

Cite Dataset ▾

Learn about [Data Citation Standards](#).

Access Dataset ▾

Contact Owner

Share

Dataset Metrics ?





15,218 Views ?

7,235 Downloads ?

3 Citations ?

Description ?

This dataset comes from the European Union project DROPS (DROught-tolerant yielding PlantS). A panel of 256 maize hybrids was grown with two water regimes (irrigated or rainfed), in seven fields in 2010 and 2012, respectively, spread along a climatic transect from western to eastern Europe, plus

<input type="checkbox"/>	 <p>10-Info-ObservedVariable.tab Tabular Data - 21.8 KB Published Nov 3, 2022 196 Downloads 17 Variables, 91 Observations UNF:6:bxOj...iaQ== </p> <p>List of phenotypic and environmental variables used in the dataset, following the MIAPPE data standard (https://www.miappe.org/)</p> <p>Observed Variable</p>
<input type="checkbox"/>	 <p>11-Info-Study.tab Tabular Data - 6.4 KB Published Nov 3, 2022 180 Downloads 11 Variables, 19 Observations UNF:6:9nO7...GEQ== </p> <p>List of studies, including locations, used in the dataset, following the MIAPPE data standard (https://www.miappe.org/)</p> <p>Study</p>

VariableID	VariableName	VariableAccessionNumber
Tnight	Night temperature	EIPO:0000001
Ri	Solar radiation	EIPO:0000002

StudyTitle	StudyUniqueID	ExperimentalSiteName	GeographicLocationLatitude	GeographicLocationLongitude
Biogemma Gaillac 2012	Gai12	Gaillac	43.9	1.89
Biogemma Gaillac 2013	Gai13	Gaillac	43.9	1.89



8-Info-BiologicalMaterial.tab

Tabular Data - 49.9 KB

Published Nov 3, 2022

159 Downloads

20 Variables, 256 Observations UNF:6:xjhU...4SA==

This file contains the description of the genotypes. Briefly, all studied hybrids result from a F1 cross

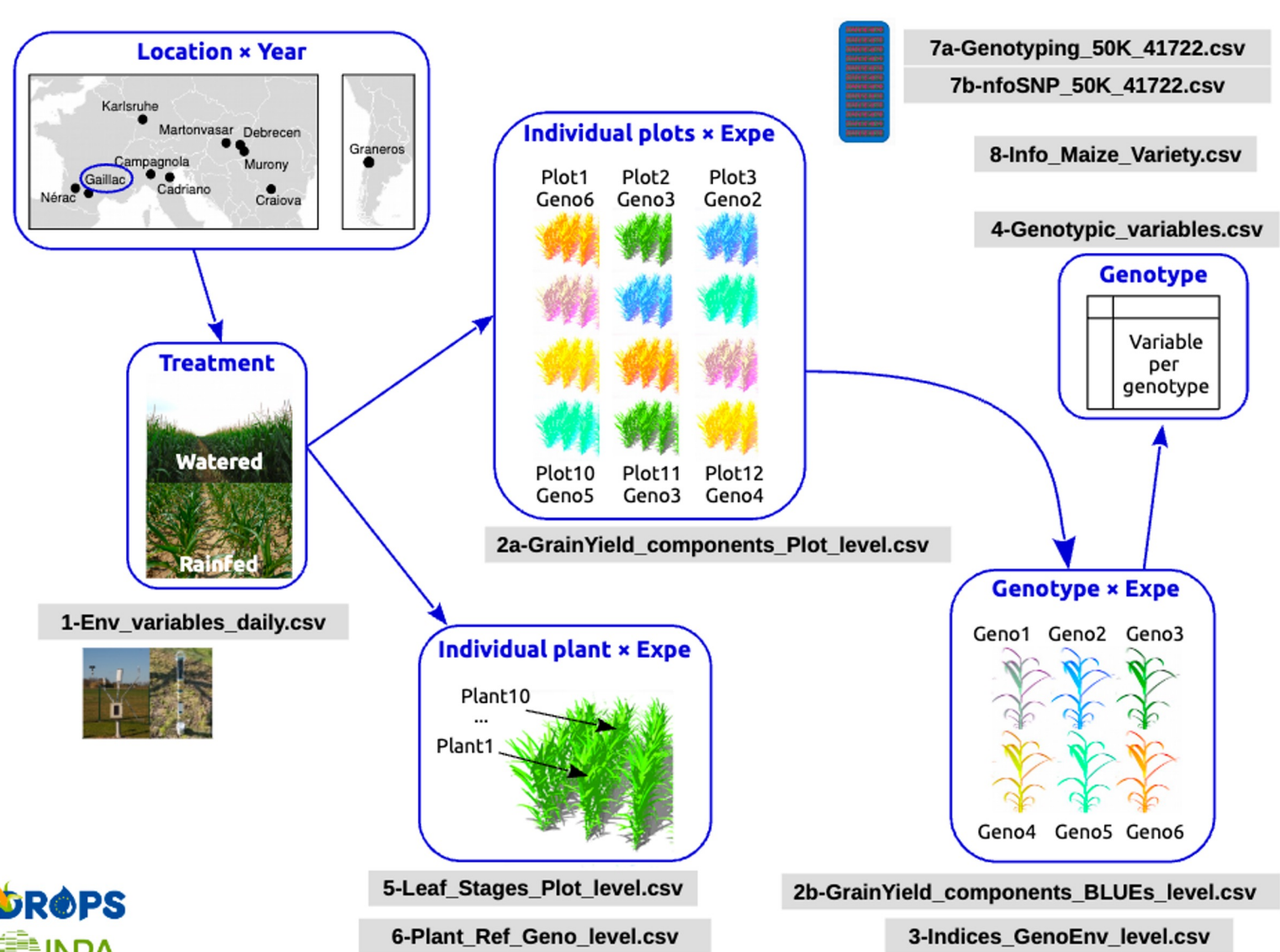
name»: Identifiers of the holding institutions. The Following columns have been added to comply to MIAPPE: "MaterialSourceDOI": DOI of the Accession identifying this variety, "MaterialSourceID": ID of the Accession identifying this variety, "BiologicalMaterialID": ID used in the data files, "Organism":NCBI Taxonomy identifier, "Genus", "Species", "InfraspecificName": variety name

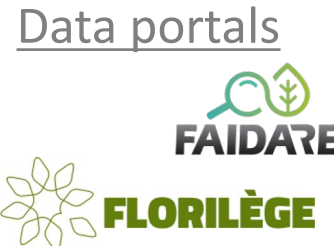
Biological Material

VarietyID	AccessionID	AccessionHolding	MaterialSourceDOI	MaterialSourceID
11430	11430_H	inra	https://doi.org/10.15454/MDSCXQ	inra:11430_H
A3	A3_H	inra	https://doi.org/10.15454/GAHOFA	inra:A3_H
A310	A310_H	inra	https://doi.org/10.15454/CNJTVT	inra:A310_H

Good practice: Human friendly Provenance

Links between files





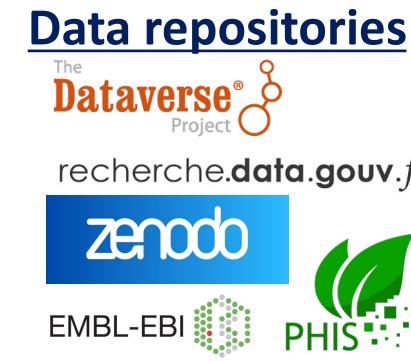
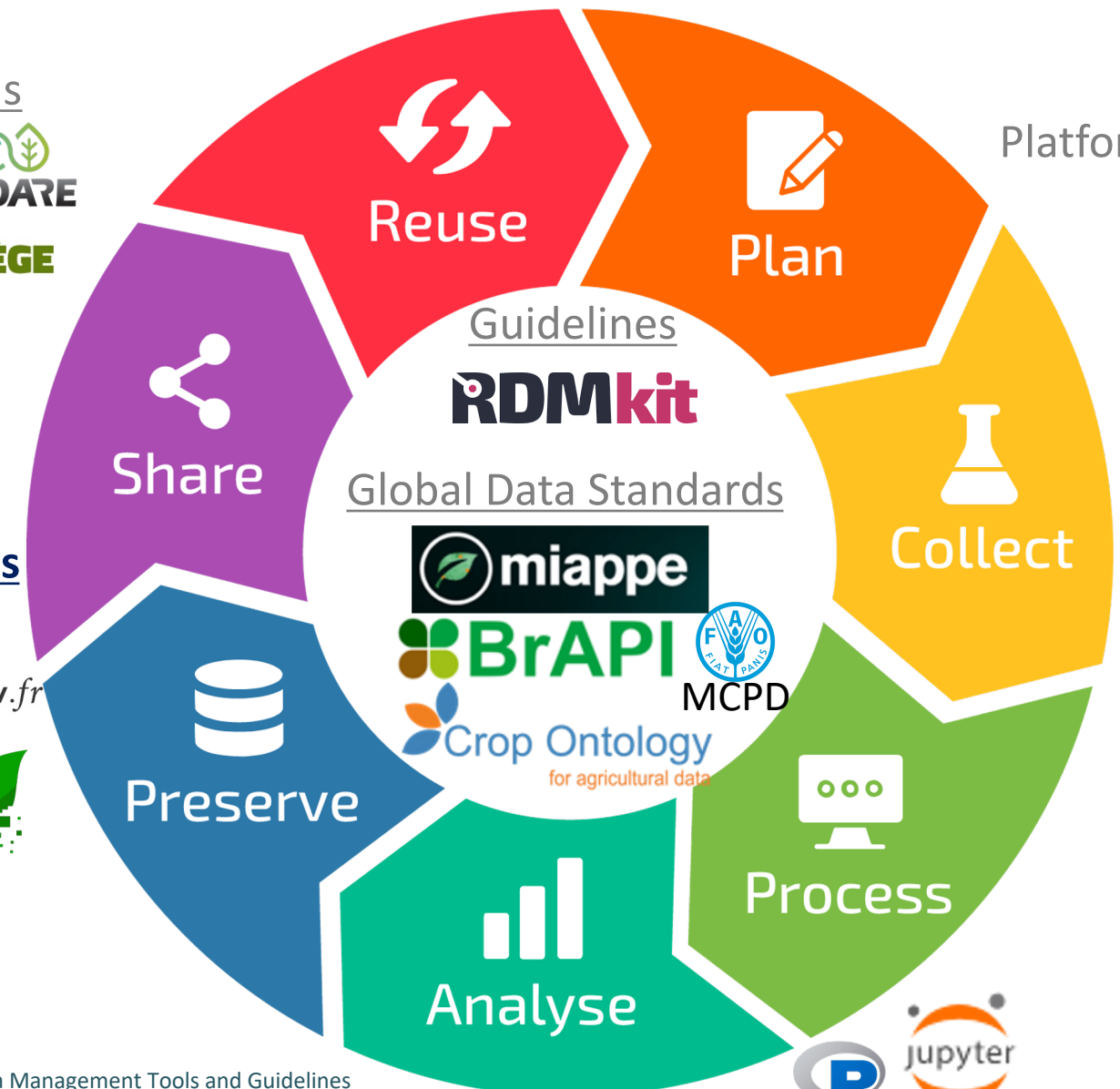
Platform Information Systems

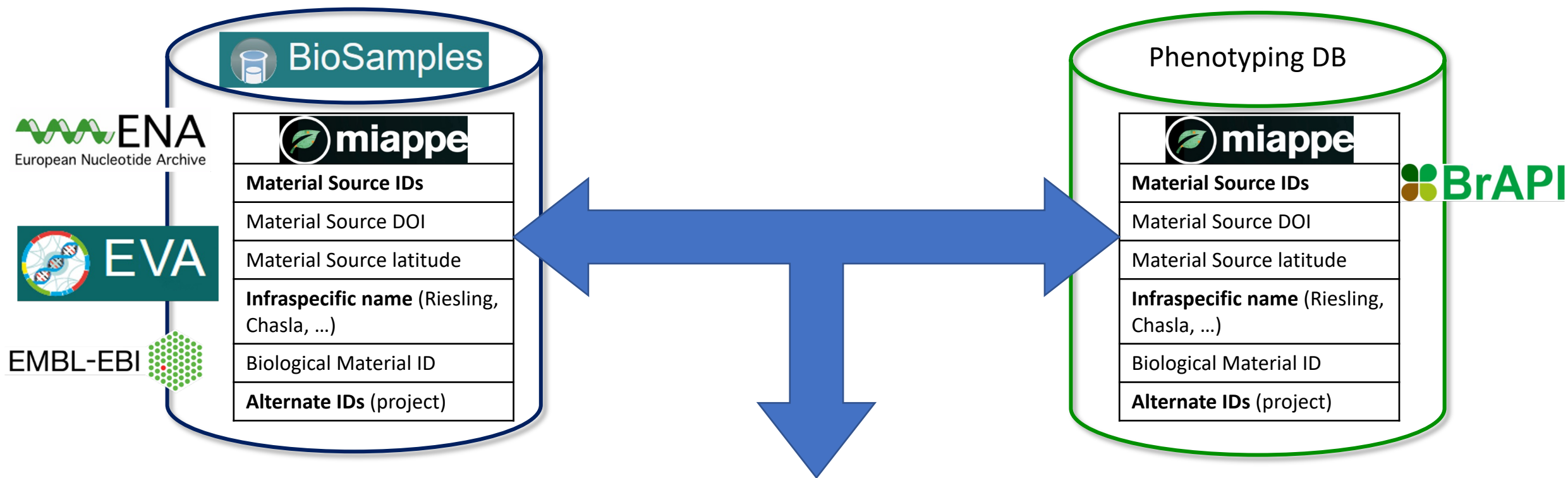


Portable devices, sensors, ...



Scripts and Workflows





Community data discovery portals

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

FAIR Data-finder for Agronomic REsearch

Search keywords

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 (1,000,000)

Germplasm Trait Reset all

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

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



Data portals



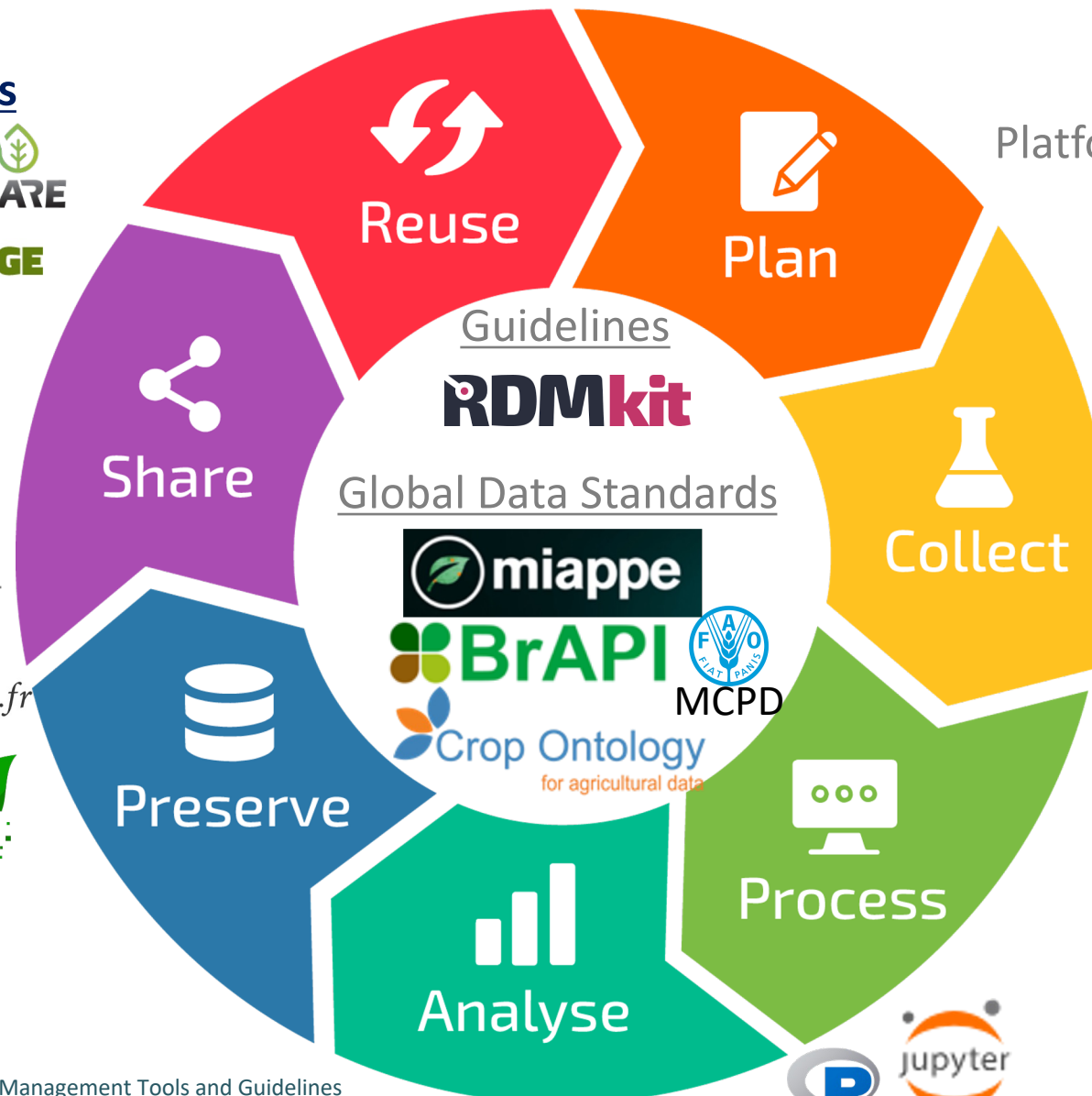
Platform Information Systems



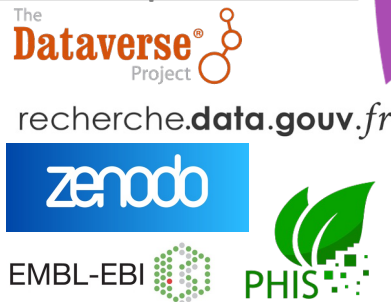
Portable devices, sensors, ...



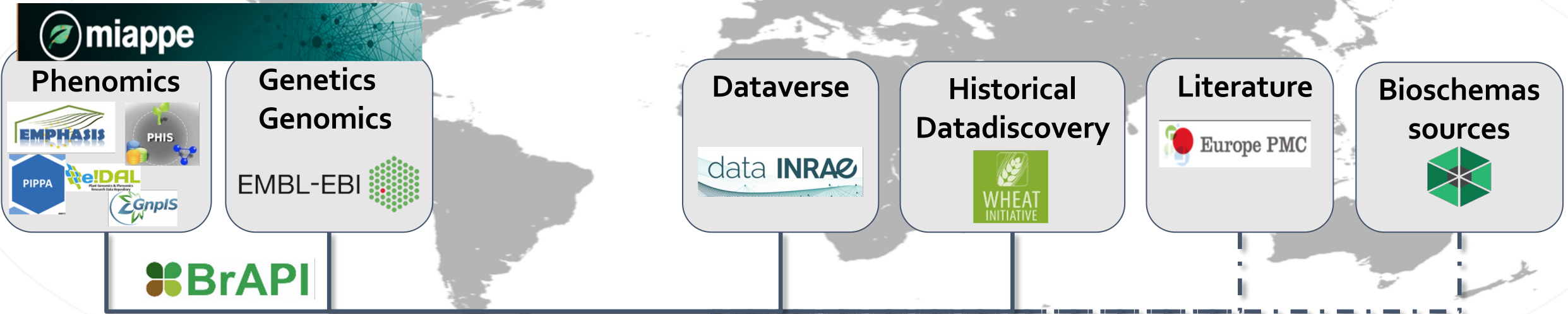
Scripts and Workflows



Data repositories



FAIDARE: Global Plant Research Data discovery portal



<https://urgi.versailles.inrae.fr/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**...

trichocarpa (poplar) located between positions 7073249 and 7073695 on chr03_scaffold_3 and which

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**
 - Budget date **Trait**
 - BS_date: Budget date **Variable**

Identifier CO_S57:1000009
 Name Budget date
 Description Assessment of the date when budget 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_S57:2000014
 Name Bud date protocol
 Description Estimated date from polynomial regression of a time series of budflush or budget scores
 Class Computation
 Calendar day **Scale**
 Identifier CO_S57:3000043
 Name Calendar day
 Data type Date
 Min 0
 Max 0
 Documentation <https://urgi.versailles.inrae.fr/>
 Context of use Research-intensive characterization
 Trial evaluation
 Breeding criterion
 Status Standard for INRAE

OK Cancel

Full text
 +
 Fine criteria
 +
Link back



- 33 databases indexed all over the world

Database (33)

Filter on Database...

Data provider

- INRAE-URGI [47,461,178]
- EBI [12,794,551]
- IPK [402,128]
- Gramene [257,561]
- T3 [223,013]
- UWA [167,167]
- PGSB [140,138]
- Rothamsted Research [137,917]
- EVA [66,510]
- GrainGenes [23,339]
- WUR [6,660]
- CIMMYT [1,788]
- TERRA-REF [1,098]
- CIRAD [722]
- NIB [694]
- VIB [692]
- IBET [67]
- DDBJ [58]
- IPGPAS [8]

France

-
-
-

Germany

-
-
-

United Kingdom

-
-
-
-

United States

-
-
-

Mexico

-

France

-
-

Europe

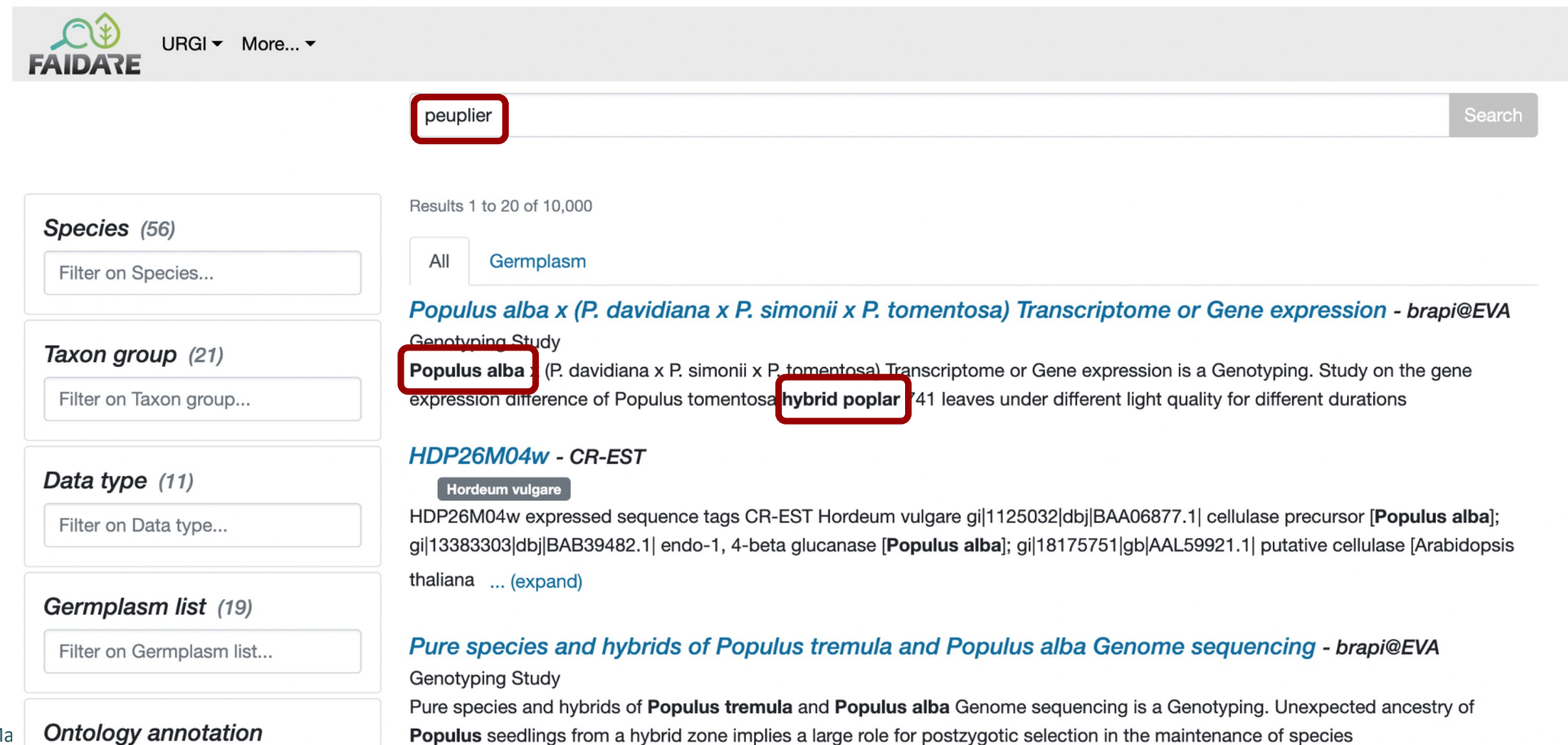
-
-
-
-
-

United States

-

Minimal generic (Data Discovery)

- 33 databases indexed all over the world
- Simple full text search
 - Ontology based synonyms



The screenshot shows the FAIDARE search interface. At the top, the FAIDARE logo and 'URGI' are visible. A search bar contains the word 'peuplier' and a 'Search' button. Below the search bar, the results are displayed as 'Results 1 to 20 of 10,000'. On the left side, there are several filter panels: 'Species (56)', 'Taxon group (21)', 'Data type (11)', 'Germplasm list (19)', and 'Ontology annotation'. Each panel has a 'Filter on...' button. The main results area shows a list of search results. The first result is titled 'Populus alba x (P. davidiana x P. simonii x P. tomentosa) Transcriptome or Gene expression - brapi@EVA'. Below the title, it says 'Genotyping Study'. The text of the result is: 'Populus alba (P. davidiana x P. simonii x P. tomentosa) Transcriptome or Gene expression is a Genotyping. Study on the gene expression difference of Populus tomentosa hybrid poplar 41 leaves under different light quality for different durations'. The words 'Populus alba' and 'hybrid poplar' are highlighted with red boxes. Below this result, there is another result titled 'HDP26M04w - CR-EST' with a 'Hordeum vulgare' tag. The text of this result is: 'HDP26M04w expressed sequence tags CR-EST Hordeum vulgare gi|1125032|dbj|BAA06877.1| cellulase precursor [Populus alba]; gi|13383303|dbj|BAB39482.1| endo-1, 4-beta glucanase [Populus alba]; gi|18175751|gb|AAL59921.1| putative cellulase [Arabidopsis thaliana ... (expand)']'. Below this, there is a third result titled 'Pure species and hybrids of Populus tremula and Populus alba Genome sequencing - brapi@EVA'. The text of this result is: 'Pure species and hybrids of Populus tremula and Populus alba Genome sequencing is a Genotyping. Unexpected ancestry of Populus seedlings from a hybrid zone implies a large role for postzygotic selection in the maintenance of species'.

- 33 databases indexed all over the world
- Simple full text search
 - Ontology based synonyms
- Dedicated filters
 - Data type
 - Taxonomy
 - Genetic resources
 - Crop Ontology
 - Gene Ontology
 - ...

Data type (37)

gen|

- Genome annotation [14,181,486]**
- Gene annotation [3,298,434]
- Gene [117,609]
- Genotyping Study [748]
- Genetic map [165]
- Genotyping Experiment [13]

Data type (37)

ph|

- Physical map feature [2,157,405]**
- Bibliography [2,988]
- Phenotyping Experiment [2,179]
- Phenotyping study [829]
- Phenotyping Study [816]
- Physical map [12]

Species (661)

Filter on Species...

Taxon group (156)

Filter on Taxon group...

- Triticum [20,150,119]**
- Hordeum [88,286]
- Brachypodium [67,367]
- Oryza [53,579]
- Sorghum [52,772]
- Setaria [48,627]
- Italica [48,625]
- Zea [46,973]

Other results are available. Refine your search.

Ontology variable selection

Filter English

LS_score Variable

- Wheat Crop Ontology **Ontology**
 - Abiotic stress **Trait class**
 - Agronomical **Trait class**
 - Grain weight **Trait**
 - Grain yield **Trait**
 - Lodging incidence **Trait**
 - LS_score: Lodging score **Variable**
 - Plant height **Trait**
 - Precocity **Trait**
 - Biotic stress **Trait class**
 - Other **Trait class**
 - Quality **Trait class**

Ontology name Wheat Crop Ontology

Identifier CO_321:1000099

Name LS_score

Synonyms Lodging score
Susceptibility to lodging

Institution INRA

Scientist Jacques Le Gouis

Date 15/06/2016

Crop Wheat

Cross reference WIPO:0000099

Lodging incidence **Trait**

Identifier CO_321:0000167

Name Lodging incidence

Description Indicates incidence of lodged plants.

Entity Plant

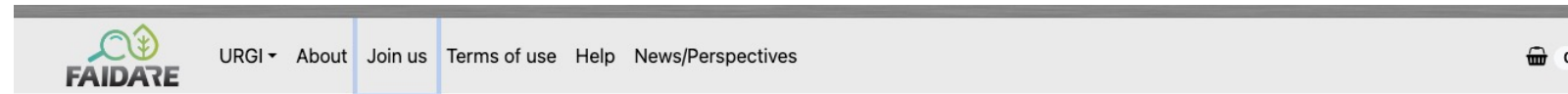
Attribute Lodging incidence

Class Agronomical

Main abbreviation Lodg

Alternative LOD

- 33 databases indexed all over the world
- Simple full text search
 - Ontology based synonyms
- Dedicated filters
 - Data type
 - Taxonomy
 - Genetic resources
 - Crop Ontology
 - Gene Ontology
 - ...
- Friendly community management
 - Join the Federation support



How to join Plant data discovery Federations (FAIDARE, WheatIS)?

Overview

The plant data discovery Federations (FAIDARE, wheatIS) provides search data portal that index the metadata from your data resources and then link back to an access page in your system. This indexation can be done using the following approaches:

- Datadiscovery files in a webfolder
- Breeding API (BrAPI) web service endpoint. Provides both datadiscovery and [summary cards](#)
- Breeding API (BrAPI) files in a webfolder. Provides both datadiscovery and [summary cards](#)

Each of those approaches are described below and all assume a minimum information set comprising an URL for link back plus description.

The metadata format must follow the indications below and we invite you to [contact us](#) as soon as possible so that we can provide help and discuss the best way to go ahead.

Breeding API (BrAPI)

This is the richer approach and will bring you all FAIDARE functionalities. The web services building will enable you to plug any [BrAPI](#) client on your database. The BrAPI file generation is simpler and easier to deploy. Only Germplasm and study are indexed from a BreedingAPI endpoint, with their full description. Those metadata will be used to create summary cards [such as](#) The datadiscovery metadata files, following the [specifications](#) below are generated from those summaries. Currently (FEB 2023), FAIDARE indexes BrAPI v1.1+ sources (V1.3 recommended).

Web services

The breedingAPI full specifications are available on www.brapi.org. The resources indexed are germplasm and study only. Information cards are created using the following calls :

- `germplasm (mandatory)`

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



URGI ▾ À propos Rejoignez-nous (EN) Conditions d'utilisation Aide Nouveautés/Perspectives Web services

Study Phenotyping Study: University_of_Bologna Cadriano 2012



📍 Origin site 📍 Collecting site 📍 Evaluation site 📍 Multi-purpose site

Identification	
Name	University_of_Bologna Cadriano 2012
Identifier	dXJuOkIOUKFFLVSR0kvc3R1ZHkvQm9sMTI=
Source	
Data link	Link to this study on URGI GnpIS

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



URGI ▾ À propos Rejoignez-nous (EN) Conditions d'utilisation Aide Nouveautés/Perspectives Web services

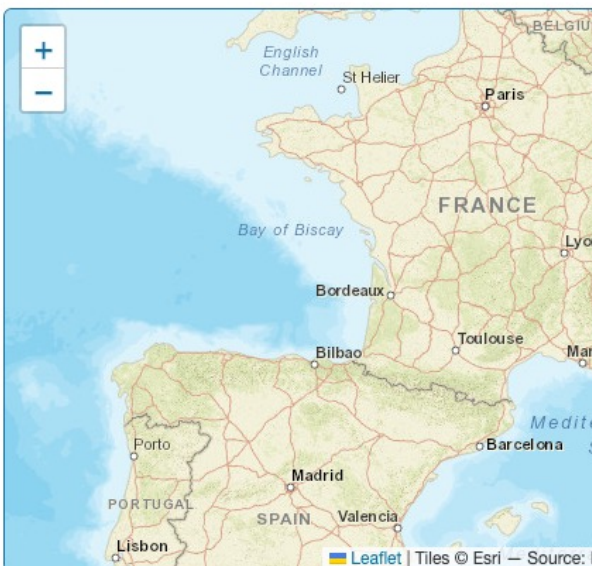
Study Phenotyping Study: University_of_Bologna Cadriano 2012

Genotype

Accession number	Name	Taxon
11430_H	11430_H	Zea may
A310_H	A310_H	Zea may
A347_H	A347_H	Zea may
A374_H	A374_H	Zea may

Variables

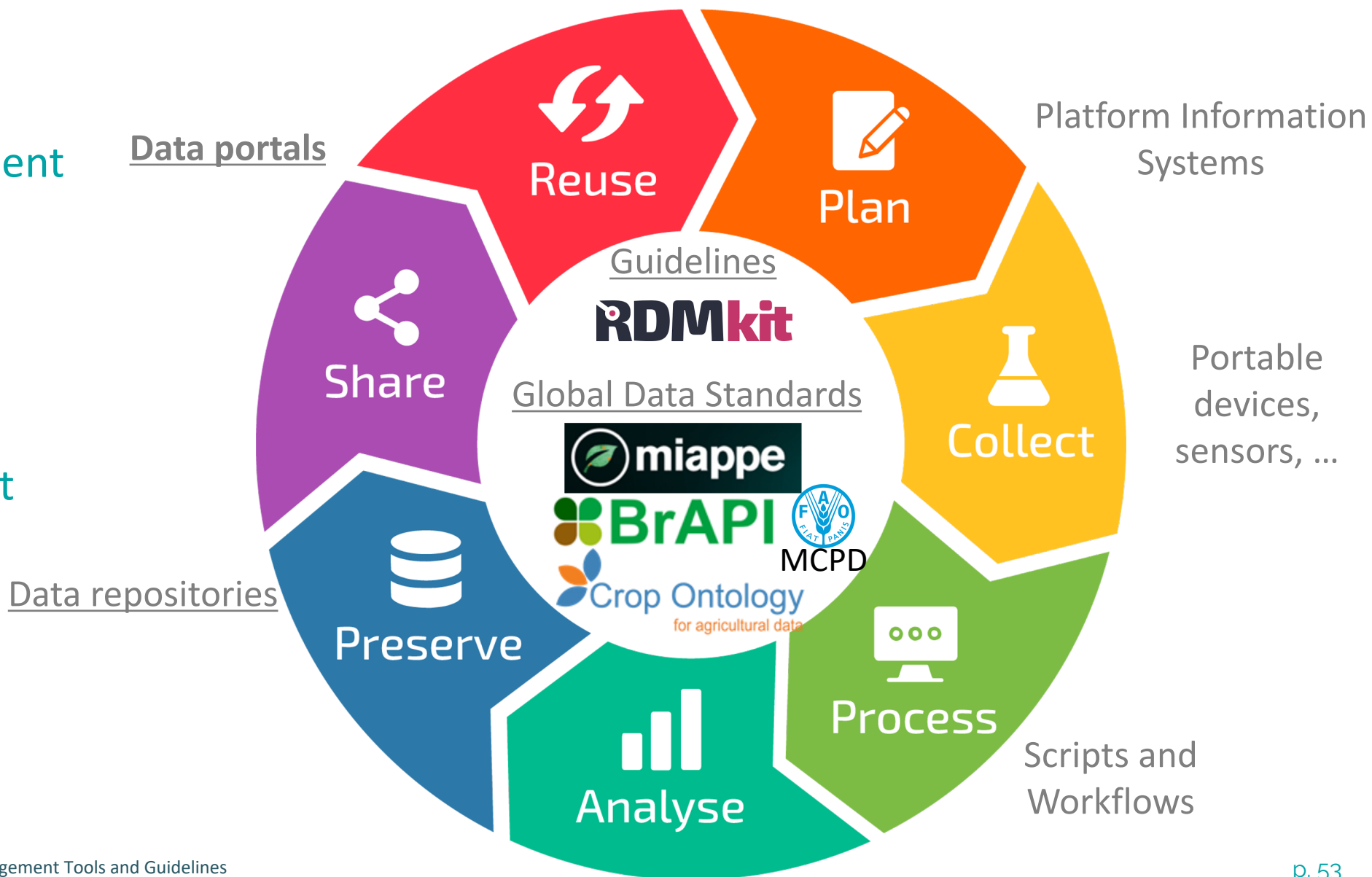
Variable ID	Variable short name	Variable long name	Ontology name	Trait description
EIPO:0000001	Tnight	Night temperature	Environmental Traits	Air night temperature
EIPO:0000002	Ri	Solar radiation	Environmental Traits	Solar radiation
EIPO:0000003	Psi	Soil water potential	Environmental	Water potential at th



Identification

Name	Univ
Identifier	dXJuOkIOUKFFLVVSR0I
Source	
Data link	Link to this study on UR

- Existing solutions and tools
- Community management
- Interoperability
 - Use cases
 - Data standards
- Global partnerships
- RDM Kit → entry point



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!

