FAIR Plant Phenomics Data Management Tools and Guidelines
Some current practices in Elixir and Emphasis European Infrastructures

www.elixir-europe.org

https://emphasis.plant-phenotyping.eu
Overview

- Data portals
  - FAIDARE
  - FLORILÈGE

- Data repositories
  - The Dataverse Project
  - recherche.data.gouv.fr
  - zenodo
  - EMBL-EBI

- Platform Information Systems
  - PHIS

- Portable devices, sensors, ...

- Scripts and Workflows

- Global Data Standards
  - miappe
  - BrAPI
  - MCPD

- RDMkit

- Guidelines

- FAIR Plant Phenomics Data Management Tools and Guidelines
  2024 Jan 13 / Cyril Pommier
Overview

- **Data portals**
  - FAIDARE
  - FLORILÈGE

- **Data repositories**
  - The Dataverse Project
  - recherche.data.gouv.fr
  - zenodo
  - EMBL-EBI
  - PHIS

- **Global Data Standards**
  - miappe
  - BrAPI
  - MCPD

- **Platform Information Systems**
  - Portable devices, sensors, ...

- **Guidelines**
  - RDMkit

- **Processes**
  - Plan
  - Collect
  - Process
  - Analyse
  - Preserve
  - Share
  - Reuse
Scope: Phenotype to genotype integration and management

- **Phenomics**
  - Dispersed (no central repo)
  - Heterogenous
  - Getting Standardized

- **Environment**
  - Healthy Soil is a Science
  - pH

- **Plant Breeding**
  - Genetic variations by Traits
  - Climate Change Studies
  - Genotype by Environment

- **Genetics Genomics Omics**
  - Mostly centralized
  - Homogenous data
  - Heterogenous metadata

- **Data Management Tools and Guidelines**
  - Findable
  - Accessible
  - Interoperable
  - Reusable
Plant data standards: involved EU research infrastructures
Gathered to solve Pheno to Omic data management and integration

• ELIXIR
  • European Infrastructure for life sciences data
  • Slovenia, France, Germany, Portugal, UK, Belgium, Italy, NI...
  • [https://elixir-europe.org/communities/plant-sciences](https://elixir-europe.org/communities/plant-sciences)
  • FAIR data management, software, training, ...

• EMPHASIS
  • European Infrastructure for Plant Phenotyping
  • France, Germany, Belgium, UK, ...
  • [https://www.plant-phenotyping.eu](https://www.plant-phenotyping.eu)
Data lifecycle: many types of phenotypes datasets

Data Acquisition

« Raw » data, pheno/env measurement, variables

- Soissons x high N
- Charger x high N
- Charger x low input
- Soissons x low input
- Soissons x high N
- Soissons x high N
- Soissons x high N
- Charger x low input

Data computation/Integration

« Derived » data, indicators

- Soissons x high N
- Charger x high N
- Soissons x low input
- Charger x low input

Genotype | Treatment | N input | Date        | Rep | Fusariose
---------|-----------|---------|-------------|-----|----------
Soissons | low input | 15,325129 | 15/11/2011  | 1   | 5         
Soissons | low input | 15,3430556 | 16/11/2011  | 2   | 7         

Genotype | Treatment | Fusariose
---------|-----------|----------
Soissons | low input | 6         

Leaf area (mm²) vs Date

FAIR Plant Phenomics Data Management Tools and Guidelines
2024 Jan 13 / Cyril Pommier
Data lifecycle: many types of phenotypes datasets

« Raw » data, pheno/env measurement, variables

- Soissons x high N
- Soissons x low input
- Charger x high N
- Charger x low input
- Charger x high N
- Charger x low input
- Soissons x high N
- Soissons x low input
- Soissons x high N
- Soissons x low input
- Soissons x high N
- Soissons x low input
- Soissons x high N
- Soissons x low input
- Soissons x high N
- Soissons x low input

Derivation, Reduction

- Soissons x high N
- Charger x high N
- Soissons x low input
- Charger x low input
- Soissons x high N
- Charger x high N
- Soissons x low input
- Charger x low input
- Soissons x high N
- Charger x high N
- Soissons x low input
- Charger x low input
- Soissons x high N
- Charger x high N
- Soissons x low input
- Charger x low input

Publication

Variety charger is resistant to fusariose under intensive cultural practice

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Treatment</th>
<th>N input</th>
<th>Date</th>
<th>Rep</th>
<th>Fusariose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soissons</td>
<td>low input</td>
<td>15,3225129</td>
<td>15/11/2011</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Soissons</td>
<td>low input</td>
<td>15,3430556</td>
<td>16/11/2011</td>
<td>2</td>
<td>7</td>
</tr>
</tbody>
</table>

Genotype | Treatment | Fusariose |
----------|-----------|-----------|
Soissons  | low input | 6         |

Data lifecycle: many types of phenotypes datasets

« Raw » data, pheno/env measurement, variables

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Treatment</th>
<th>N input</th>
<th>Date</th>
<th>Rep</th>
<th>Fusariose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soissons</td>
<td>low input</td>
<td>15,3225129</td>
<td>15/11/2011</td>
<td>1</td>
<td>5</td>
</tr>
<tr>
<td>Soissons</td>
<td>low input</td>
<td>15,3430556</td>
<td>16/11/2011</td>
<td>2</td>
<td>7</td>
</tr>
</tbody>
</table>

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

« Derived » data, indicators

<table>
<thead>
<tr>
<th>Genotype</th>
<th>Treatment</th>
<th>Fusariose</th>
</tr>
</thead>
<tbody>
<tr>
<td>Soissons</td>
<td>low input</td>
<td>6</td>
</tr>
</tbody>
</table>

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

Raw data long term conservation
Plant Data standards

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

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

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

Technical
- Data integration and sharing
- Interoperability: tools and systems
  - GA4GH
  - Breeding API www.brapi.org
  - Computer scientist driven
• **Phenotype **Structure **Standard**

Minimal Information About Plant Phenotyping Experiments

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

Version 1
- 2015
  - Initial Checklist

Version 1.1
- 2019
  - Data model formalisation

Version 1.2
- 2024
  - Minor additions
  - RFC opens in January ➔ shared through the mailing list
• 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 Selby)

  - **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

Enabling reusability of plant phenomic datasets with MIAPPE 1.1


### MIAPPE Specifications

- **Specification table**
- **Sections**
- **Metadata Fields**

<table>
<thead>
<tr>
<th>line #</th>
<th>MIAPPE Check list</th>
<th>Definition</th>
<th>MIAPPE</th>
<th>Example</th>
<th>Format</th>
<th>Cardinality</th>
</tr>
</thead>
<tbody>
<tr>
<td>DM-1</td>
<td>Investigation</td>
<td>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).</td>
<td></td>
<td></td>
<td></td>
<td>1 per MIAPPE submission</td>
</tr>
<tr>
<td>DM-10</td>
<td>Study</td>
<td>A study (or experiment) comprises a series of assays (or measurements) of one or more types, undertaken to answer a particular biological question.</td>
<td></td>
<td></td>
<td></td>
<td>1+ per investigation</td>
</tr>
<tr>
<td>DM-30</td>
<td>Person</td>
<td>A human involved in the investigation or specifically any of its studies.</td>
<td></td>
<td></td>
<td></td>
<td>1+ per investigation / observation</td>
</tr>
<tr>
<td>DM-36</td>
<td>Data File</td>
<td>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.</td>
<td></td>
<td></td>
<td></td>
<td>0+ per study</td>
</tr>
<tr>
<td>DM-40</td>
<td>Biological Material</td>
<td>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 clone) is called the material source, which, when held by a material repository, should have its stock-identified.</td>
<td></td>
<td></td>
<td></td>
<td>1+ per study; 0+ per observation unit</td>
</tr>
<tr>
<td>DM-57</td>
<td>Environment</td>
<td>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).</td>
<td></td>
<td></td>
<td></td>
<td>0-1 per study</td>
</tr>
<tr>
<td>DM-60</td>
<td>Experimental Factor</td>
<td>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)</td>
<td></td>
<td></td>
<td></td>
<td>0+ per study; 0+ per observation unit</td>
</tr>
<tr>
<td>DM-64</td>
<td>Event</td>
<td>An event is a 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 a subset of observation units.</td>
<td></td>
<td></td>
<td></td>
<td>0+ per study/observation</td>
</tr>
<tr>
<td>DM-69</td>
<td>Observation Unit</td>
<td>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.</td>
<td></td>
<td></td>
<td></td>
<td>1+ per study</td>
</tr>
<tr>
<td>DM-75</td>
<td>Sample</td>
<td>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).</td>
<td></td>
<td></td>
<td></td>
<td>0+ per observation unit</td>
</tr>
<tr>
<td>DM-82</td>
<td>Observed Variable</td>
<td>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.</td>
<td></td>
<td></td>
<td></td>
<td>1+ per study</td>
</tr>
<tr>
<td>DM-83</td>
<td>Variable ID</td>
<td>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: &lt;trait abbreviation&gt;<em>&lt;method abbreviation&gt;</em>&lt;scale abbreviation&gt;. A variable ID must be unique within a given investigation.</td>
<td>Ant_Cmp_Cday</td>
<td>Unique identifier</td>
<td></td>
<td>1</td>
</tr>
<tr>
<td>DM-84</td>
<td>Variable name</td>
<td>Name of the variable.</td>
<td>Anthesis computed in growing degree days</td>
<td>Free text</td>
<td></td>
<td>0-1</td>
</tr>
<tr>
<td>DM-85</td>
<td>Variable accession number</td>
<td>Accession number of the variable in the Crop Ontology</td>
<td>CO_322_0000794</td>
<td>Crop Ontology term</td>
<td>0-1</td>
<td></td>
</tr>
<tr>
<td>DM-86</td>
<td>Trait</td>
<td>Name of the (plant or environmental) trait under observation</td>
<td>Anthesis time</td>
<td>Reproductive growth time</td>
<td>Free text</td>
<td>1</td>
</tr>
<tr>
<td>DM-87</td>
<td>Trait accession number</td>
<td>Accession number of the trait in a suitable controlled vocabulary (Crop Ontology, Trait Ontology).</td>
<td>CO_322_0000030</td>
<td>Term from Plant Trait Ontology, Crop Ontology, or XML Environment Ontology</td>
<td>0-1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>MIAPPE Data Management Tools and Guidelines 2024 Jan 13 / Cyril Pommier</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
- MIAPPE Specifications
- Specification table
- Sections
- Metadata Fields
- Linked and organized
• **Investigation**: whole dataset
• **Study**: one experiment in one location for one to several years
• **Assay**:
  • trait or indice (Pheno or Env) observed
  • Level + Trait + Method + Scale/Unit

• Level:
  • Plant
  • Microplot
  • Block
  • Trial
  • ...

MIAPPE V1.1 Overview the (ISA) backbone

Investigation

Study

Assay (Observed variable)

Data file
MIAPPE main sections – Biological material

• Plant Material
  ◦ Identification
  ◦ Description

• Multi Crop Passport Descriptor

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, ...

MCPD identification system:
  • Genebank/Lab + Species + accession number
  • DOI

• Lab + internal accession number
• URI

• Lab + internal accession number (mandatory)
• BioSample ID
MIAPPE main sections – Observed Variable (Assay)

Plant Phenomic Specific Ontology Model:

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

- T1: Plant Height
- M1: Total height
- M2: First tassel branch
- M3: Last expanded leaf
- M4: Youngest growing leaf
- M5: Highest pixel corresponding to plant

- U1: cm
- U2: mm
- U3: pixel
• MIAPPE – Other Important sections
• MIAPPE Overview
  Data file content

• Any format (Near Infra Red Spectrum, Images, Image Archives references, ....)

• Mostly tabular

<table>
<thead>
<tr>
<th>A</th>
<th>B</th>
<th>C</th>
<th>D</th>
<th>E</th>
<th>F</th>
<th>G</th>
<th>H</th>
<th>I</th>
</tr>
</thead>
<tbody>
<tr>
<td>661300270 Ardon</td>
<td>2004</td>
<td>45.645632645603683</td>
<td>12/01/2004</td>
<td>284.3</td>
<td>12/01/2004</td>
<td>14.630625</td>
<td>12/05/2005</td>
<td></td>
</tr>
<tr>
<td>661300270 Ardon</td>
<td>2005</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>661300444 Ardon</td>
<td>2004</td>
<td>38.96112577281653</td>
<td>12/01/2004</td>
<td>228.8</td>
<td>12/01/2004</td>
<td>8.5030559999999991</td>
<td>12/05/2005</td>
<td></td>
</tr>
<tr>
<td>661300444 Ardon</td>
<td>2005</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>661300312 Cavalleri Maggiore</td>
<td>2004</td>
<td>52.4</td>
<td>01/01/2004</td>
<td>249.9</td>
<td>01/01/2004</td>
<td>12.981609000000001</td>
<td>01/05/2005</td>
<td></td>
</tr>
<tr>
<td>661300312 Cavalleri Maggiore</td>
<td>2005</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>661300371 Cavalleri Maggiore</td>
<td>2004</td>
<td>45.74</td>
<td>01/01/2004</td>
<td>230.2</td>
<td>01/01/2004</td>
<td>10.3041</td>
<td>01/05/2005</td>
<td></td>
</tr>
<tr>
<td>661300371 Cavalleri Maggiore</td>
<td>2005</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>661300487 Cavalleri Maggiore</td>
<td>2004</td>
<td>72.52</td>
<td>01/01/2004</td>
<td>309.8</td>
<td>01/01/2004</td>
<td>10.679823999999998</td>
<td>01/05/2005</td>
<td></td>
</tr>
<tr>
<td>661300487 Cavalleri Maggiore</td>
<td>2005</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>661300585 Cavalleri Maggiore</td>
<td>2004</td>
<td>71.739999999999995</td>
<td>01/01/2004</td>
<td>305.7</td>
<td>01/01/2004</td>
<td>10.956100000000001</td>
<td>01/05/2005</td>
<td></td>
</tr>
<tr>
<td>661300585 Cavalleri Maggiore</td>
<td>2005</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>661300468 Headley</td>
<td>2004</td>
<td>45.27</td>
<td>01/01/2004</td>
<td>247</td>
<td>01/01/2004</td>
<td>15.888196000000002</td>
<td>01/05/2005</td>
<td></td>
</tr>
<tr>
<td>661300468 Headley</td>
<td>2005</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>661300469 Headley</td>
<td>2004</td>
<td>70.930000000000007</td>
<td>01/01/2004</td>
<td>313</td>
<td>01/01/2004</td>
<td>13.271448999999999</td>
<td>01/05/2005</td>
<td></td>
</tr>
<tr>
<td>661300469 Headley</td>
<td>2005</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>661300533 Headley</td>
<td>2004</td>
<td>57.67</td>
<td>01/01/2004</td>
<td>258.8</td>
<td>01/01/2004</td>
<td>01/01/2004</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
MIAPE perspectives

- Recommended data file format
- Extend standard interoperability toward ICASA
- Github issues
  - [https://github.com/MIAPPE/MIAPPE/issues](https://github.com/MIAPPE/MIAPPE/issues)
Overview

- Data portals
  - FAIDARE
  - FLORILÈGE

- Data repositories
  - The Dataverse Project
  - recherche.data.gouv.fr
  - Zenodo
  - EMBL-EBI

- Global Data Standards
  - miappe

- Platform Information Systems
  - PHIS

- Portable devices, sensors, ...

- Scripts and Workflows

- Reuse
- Plan
- Collect
- Process
- Analyse
- Preserve
- Share

Guidelines

RDMkit

MCPD

Crop Ontology

for agricultural data
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

- **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
- 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
Your domain

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 MIAPE standard?
- Are you exchanging data with individual researchers?
  - What media is data being collected?
  - Is the data described in a MIAPE-compliant manner?
- Are you exchanging data across different data management platforms?
  - Do these platforms implement the Breeding API BrAPI specification?
  - If not, are they MIAPE-compliant and do they enable automated data exchange?

Solutions

Checklists and ontologies

- The metadata standard applicable to plant phenotyping experiments is MIAPE.
- There is a section dedicated to the identification of plant biological materials that follows Multi-Crop Passport Descriptor (MCMD) 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)
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 infraspecific 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...
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.

COVID-19 Data Portal

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

CSC

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

IFB

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

Plant phenomics

Plant genomics

Figure 1. The plant phenomics tool assembly.

Figure 1. The plant genomics tool assembly.
Tool assemblies for plant data

Plant phenomics

Plant genomics
Plant Genomic assembly

Publish Sequence and Variation @ EMBL-EBI

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

[Diagram]

https://rdmkit.elixir-europe.org/plant_genomics_assembly
Plant Phenomic 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 MIAPPE

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

Figure 1. The plant phenomics tool assembly.
Dataverse


https://demo.dataverse.org/
Step 1: Dataset creation: Minimal metadata

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 the community.
  - Examples: Data INRAE and CIRAD.

- **Jülich DATA** (NRW - DE)
  - Open to submission from any research activity done from partners at Forschungszentrum Jülich.
  - Meant for data and software submissions.
  - Maintained by the central library of FZJ.
  - Examples under subject “Agricultural Sciences”.

- **dmportal.biodata.pt** (PT)
  - Open to submission of biological data from Portuguese research & innovation communities.
  - Example: Plant BioDataVerse.
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
  - "Intraspecific 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](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**

Millet, Emilie J.; Pommier, Cyril; Buy, Mélanie; Nagel, Axel; Kruijer, Willem; Welz-Bolduan, Therese; Lopez, Jeremy; Richard, Cécile; Racz, Ferenc; Tanzi, Franco; Spitkot, Tamás; Canè, Maria-Angela; Negro, Sandra S.; Coupled-Ledru, Aude; Nicolas, Stéphane D.; Palaffre, Carine; Bauland, Cyril; Praud, Sébastien; Ranc, Nicolas; Prester, 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](https://doi.org/10.15454/IASSTN), Recherche Data Gouv, V4, UNF:6:zS2/ccOQxFrKiU+1S0Cvg== [fileUNF]

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 2016 and 2017, respectively, spread along a climatic transect from western to eastern Europe.
### MIAPPE Metadata

#### 10-Info-ObservedVariable.tab
- Tabular Data: 21.8 KB
- Published Nov 3, 2022
- 196 Downloads
- 17 Variables, 91 Observations
- List of phenotypic and environmental variables used in the dataset, following the MIAPPE data standard (https://www.miappe.org/)

<table>
<thead>
<tr>
<th>VariableID</th>
<th>VariableName</th>
<th>VariableAccessionNumber</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tnnight</td>
<td>Night temperature</td>
<td>EIPO:00000001</td>
</tr>
<tr>
<td>Ri</td>
<td>Solar radiation</td>
<td>EIPO:00000002</td>
</tr>
</tbody>
</table>

#### 11-Info-Study.tab
- Tabular Data: 6.4 KB
- Published Nov 3, 2022
- 180 Downloads
- 11 Variables, 19 Observations
- List of studies, including locations, used in the dataset, following the MIAPPE data standard (https://www.miappe.org/)

<table>
<thead>
<tr>
<th>StudyTitle</th>
<th>StudyUniqueID</th>
<th>ExperimentalSiteName</th>
<th>GeographicLocationLatitude</th>
<th>GeographicLocationLongitude</th>
</tr>
</thead>
<tbody>
<tr>
<td>Biogemma Gaillac 2012</td>
<td>Gai12</td>
<td>Gaillac</td>
<td>43.9</td>
<td>1.89</td>
</tr>
<tr>
<td>Biogemma Gaillac 2013</td>
<td>Gai13</td>
<td>Gaillac</td>
<td>43.9</td>
<td>1.89</td>
</tr>
</tbody>
</table>
MIAPPE Metadata

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 between varieties of different species. The following information is included:

- **name**: Identifiers of the holing 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

<table>
<thead>
<tr>
<th>VarietyID</th>
<th>AccessionID</th>
<th>AccessionHolding</th>
<th>MaterialSourceDOI</th>
<th>MaterialSourceID</th>
</tr>
</thead>
<tbody>
<tr>
<td>11430</td>
<td>11430_H</td>
<td>inra</td>
<td><a href="https://doi.org/10.15454/MDSCXQ">https://doi.org/10.15454/MDSCXQ</a></td>
<td>inra:11430_H</td>
</tr>
<tr>
<td>A3</td>
<td>A3_H</td>
<td>inra</td>
<td><a href="https://doi.org/10.15454/GAHOFA">https://doi.org/10.15454/GAHOFA</a></td>
<td>inra:A3_H</td>
</tr>
<tr>
<td>A310</td>
<td>A310_H</td>
<td>inra</td>
<td><a href="https://doi.org/10.15454/CNJTVT">https://doi.org/10.15454/CNJTVT</a></td>
<td>inra:A310_H</td>
</tr>
</tbody>
</table>
Good practice: Human friendly Provenance

Links between files
Overview

Guidelines

Global Data Standards

Data portals

Data repositories

Data Management Tools and Guidelines

Portable devices, sensors, ...

Platform Information Systems

Scripts and Workflows

Reuse

Plan

Collect

Preserve

Process

Analyse

Share
Data Integration between silos, From Phenotyping to Genotyping

- **BioSamples**
  - Material Source IDs
  - Material Source DOI
  - Material Source latitude
  - Infraspecific name (Riesling, Chasla, ...)
  - Biological Material ID
  - Alternate IDs (project)

- **Phenotyping DB**
  - Material Source IDs
  - Material Source DOI
  - Material Source latitude
  - Infraspecific name (Riesling, Chasla, ...)
  - Biological Material ID
  - Alternate IDs (project)

Community data discovery portals

**FAIR Data-finder for Agronomic REsearch**

- Sources
  - URGI GenPls (81,335)
  - EBI European Nucleotide Archive (44,975)
  - CIRAD TropGENE (722)
  - VIB PIPPA (692)
  - IBET BioDate (67)
  - NGS@GenPls (78,814,632)
  - EvolTree@GenPls (5,354)
  - OpenMinTeD@GenPls (3,392)
  - EBI Ensemble Plants

- Crops
  - Search crops

- Germplasm list
  - Search germplasm lists
Overview

Data portals
- FAIDARE
- FLORILÈGE

Data repositories
- The Dataverse Project
- recherche.data.gouv.fr
- zenodo
- EMBL-EBI
- PHIS

Data management tools
- RDMkit
- BrAPI
- miappp
- Crop Ontology

Guidelines
- MCPD
- Guidelines

Global Data Standards

Platform Information Systems

Portable devices, sensors, ...

Scripts and Workflows

FAIR Plant Phenomics Data Management Tools and Guidelines
2024 Jan 13 / Cyril Pommier
FAIDARE: Global Plant Research Data discovery portal

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

Full text +
Fine criteria +
Link back
• 33 databases indexed all over the world
FAIDARE: Global Plant Research Data discovery portal

- 33 databases indexed all over the world
- Simple full text search
  - Ontology based synonyms
• 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
  • …
• 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 Germplams 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 germplasms and study only. Information cards are created using the following calls:
- germplasms (mandatory)
Study Phenotyping Study: University_of_Bologna Cadriano 2012

Identification

<table>
<thead>
<tr>
<th>Name</th>
<th>University_of_Bologna Cadriano 2012</th>
</tr>
</thead>
<tbody>
<tr>
<td>Identifier</td>
<td>dXJuOkiOUkFILVSR0kvc3R1ZHk=Qm9sMTI=</td>
</tr>
<tr>
<td>Source</td>
<td><a href="https://urgi.versailles.inrae.fr/gnpis/">GnpiS</a></td>
</tr>
<tr>
<td>Data link</td>
<td>Link to this study on URGi GnpiS</td>
</tr>
</tbody>
</table>
Study Phenotyping Study: University_of_Bologna_Cadriano
2012

Genotype

<table>
<thead>
<tr>
<th>Accession number</th>
<th>Name</th>
<th>Taxon</th>
</tr>
</thead>
<tbody>
<tr>
<td>11430_H</td>
<td>11430_H</td>
<td>Zea mays</td>
</tr>
<tr>
<td>A310_H</td>
<td>A310_H</td>
<td>Zea mays</td>
</tr>
<tr>
<td>A347_H</td>
<td>A347_H</td>
<td>Zea mays</td>
</tr>
<tr>
<td>A374_H</td>
<td>A374_H</td>
<td>Zea mays</td>
</tr>
</tbody>
</table>

Variables

<table>
<thead>
<tr>
<th>Variable ID</th>
<th>Variable short name</th>
<th>Variable long name</th>
<th>Ontology name</th>
<th>Trait description</th>
</tr>
</thead>
<tbody>
<tr>
<td>EIP0:0000001</td>
<td>Tnight</td>
<td>Night temperature</td>
<td>Environmental</td>
<td>Air night temperature</td>
</tr>
<tr>
<td>EIP0:0000002</td>
<td>Ri</td>
<td>Solar radiation</td>
<td>Environmental</td>
<td>Solar radiation</td>
</tr>
<tr>
<td>EIP0:0000003</td>
<td>Psi</td>
<td>Soil water potential</td>
<td>Environmental</td>
<td>Water potential at th...</td>
</tr>
</tbody>
</table>
Take home message

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

Data portals

Reuse
Plan
Guidelines
Global Data Standards

Collect
Portable devices, sensors, ...

Preserve
Data repositories

Process
Scripts and Workflows

Analyse

Miappe

BrAPI

MCPD

Global Partnerships

RDM Kit à entry point

Platform Information Systems
Aknowledgments

Elixir Plant community & platforms
Beier S., Gruden C., Pommier C., Coppens F, Scholz L., Lange M., Contreras B., AdamBlondon 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,
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!

Acknowledgments

Elixir Plant community & platforms
Beier S., Gruden C., Pommier C., Coppens F., Scholz L., Lange M., Contreras B., AdamBlondon 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,
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!

Acknowledgments

Elixir Plant community & platforms
Beier S., Gruden C., Pommier C., Coppens F., Scholz L., Lange M., Contreras B., AdamBlondon 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,
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!