Data Management

S5: Advanced data management (2)

Towards Data Science

Monday 11/12/2023 - 14:00-16:00 (CET)

Isabelle Alic, Farzaneh Kazemipour-Ricci - INRAE
Marie Weiss, Llorenç Cabrera Bosquet - INRAE
General objectives: Overview of data management for plant phenotyping - focus on FAIR data

Session 5

Advanced data management (2) - Towards Data Science
Overview
Towards Data Science

• Introduction & Review of last sessions
• Use case: from raw data to process analyses based on FAIR principles - 4P
• Data enrichment: Events, Annotation and Documents
• Use Case: Data analysis and integration
• Geospatial: Integration and a step towards decision support
• Conclusion & Perspectives
Introduction & Review of last sessions

Reminder - Data Management Best Practices

Step 1 - Identification

URI of Experiment: <emphasis:study56>

URI of Facility: <emphasis:gantry2>

URI of Device: <emphasis:st14>

URI of microplot: <emphasis:plot15879>

URI of germplasm:
<https://www.geves.fr/catalogue/variete/1002491-aubusson>
Introduction & Review of last sessions

Reminder - Data Management Best Practices

Step 2 - Experiment Description

- Administrative Data
  - Description
  - Institution
  - Project
  - Facility
  - Persons

- Implementation Data
  - Factors
  - Genotype
  - Crop Management Events
  - Measured units

- MetaData
  - Green house
  - Growth chamber
  - Field

- Metabolomics Data
  - Nitrogen level
  - Watering rate
  - Organic fertilization level
  - Pesticide dose

- Tillage
  - Sowing/Harvest
  - Irrigation
  - Accidents

- Plot/ Plant/ Organ Position
Introduction & Review of last sessions

Reminder - Data Management Best Practices

Step3 - Datafiles storage

Study the storage solution best suited to your needs and uses!

Single Server storage

On Filesystem

Hosting

AND

OR

Distributed storage

NAS, File server, etc.

Cloud

Researchers including scientists, students, lecturers, teachers and citizen scientists

Explore and Contribute

Discover Research Outputs
Find datasets, scientific publications and software for your research activities

Publish Research Outputs
Store, backup, archive your data, publications, software

Tools

Access Computing and Storage Resources
Find HPC, IT centres for science, cloud computing, online storage

Process and Analyse
Verify, organise, transform and integrate data, then export it in the format you need

Find Funding Opportunities
Learn about H2020 - EU research calls

Access Training Materials
Find lecture, courses, videos, articles

More

Research Data Management
Research Infrastructures
Instruments & Equipments
Regional & Thematic Projects

Get Inspired
Introduction & Review of last sessions

Reminder - Data Management Best Practices

Step 3 - Data Declaration

- Variable description
  - Using Model
  - With ID
- Provenance
  - Data Acquisition description
  - With ID

PROVENANCE
- Description
- Start / End Date
- Vector
- Sensor
- Operator
- …
Introduction & Review of last sessions

Reminder - Data Management Best Practices

Step 3 - Data Declaration

Environmental Data

Provenance
- Devices: emphasis:st14
- Station: campbell-2018-dl53

Data
- Variable: air_temperature_minimumDaily_degC
- Date: 2023-07-01
- Target: plot2023-BH53
- Value: 12

Phenotypic Data

Provenance
- Vector: fieldRobot1
- Camera: campbell-2019-dl4
- Software: 4P
- ProvUsed: Datafile568

Data
- Variable: Canopy_CoverFraction0deg_ImageSeg_Unitless
- Date: 2023-07-01
- Target: <emphasis:plot15879>
- Value: 0.1
Introduction & Review of last sessions

Reminder - Data Management Best Practices

=> Having well-organized and well-described data will enable **data exploration and analysis**
Use case: from raw data to process analyses based on FAIR principles

4P: Plant Phenotyping Processing Platform - Marie Weiss

Field Acquisitions

Data Processing

Traits

- LIDAR
- RGB
- Multispectral
- Stereo vision

PHIS

4P

CSV

EMPHASIS

EUROPEAN INFRASTRUCTURE FOR PLANT PHENOTYPING
Survey among phenotyping installations and users @ INRAE (2023)

Prioritization of needs
1. Data traceability - FAIR: 28% Data only 71% Data+ Algorithms
2. Ergonomy / Easy to use
3. Quick Processing & computing facilities
4. Data Storage
5. Processing traceability

 Capacities for processing
1. 57%: No in-house high-speed processing capacity
2. 21%: Processing facilities within the installation/lab
3. 16%: subcontracting specialized companies
4. 6%: Not yet defined

Current Data Management
500 Go to 10 To / campaign
Requirements

• Usable by all phenotyping installations @ INRAE
  • Compatible with all vectors/sensors
  • Processing pipelines easily shareable with PHENOME partners
  • Ergonomic: users are not developers

• Traceable
  • Storage of information related to sensor/vectors/data
  • Management and versioning of data processing chains

• Flexibility
  • Genericity: able to integrate processing chains developed in different languages, use commercial softwares (e.g. metashape)
  • Management and versioning of data processing chains
  • Scalability: update with new sensors/vectors/algo/traits
Proposed Solution

Modules (Source Code & Docker)

Source Files

Groovy File
data/test

Virtual Machines

Import Export

PROCESSING ENGINE

Workflow Management Module sequencing

Cromwell

IHM

Automation Server Build/test docker

Orchestration Management Tool

Virtual Machines
Dashboard of Toulouse

- Amount of data: 11.4 TB
- Running processes: 3
- Disk space: 35%

Running processes over the last 12 months

Processes status
- Inactive
- Running
- Complete
- Failed
- Cancelled

Macros status
- Archived
- Valid
- Invalid
- Published

Datatypes
- L0 Data
- L1 Data
- L2 Data
# Data Management

<table>
<thead>
<tr>
<th>Measuring session</th>
<th>Loading date</th>
<th>Experiment(s)</th>
<th>Vector</th>
<th>Data</th>
<th>PHIS</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>DD/MM/YYYY - D</td>
<td>DD/MM/YYYY</td>
<td>All</td>
<td>All</td>
<td>All</td>
<td>All</td>
<td>All</td>
</tr>
<tr>
<td>30/03/2023 11:47</td>
<td>07/12/2023 12:08</td>
<td>23ZM8_FFAST</td>
<td>Phenomobile</td>
<td>HDF5 Files</td>
<td>RGB Images</td>
<td>Point Clouds</td>
</tr>
<tr>
<td>17/11/2023 11:34</td>
<td>04/12/2023 13:04</td>
<td>24ZN8_FFAST</td>
<td>Phenomobile</td>
<td>HDF5 Files</td>
<td>RGB Images</td>
<td>50.2 GB</td>
</tr>
<tr>
<td>14/11/2023 12:34</td>
<td>27/11/2023 13:29</td>
<td>24ZN8_FFAST</td>
<td>Phenomobile</td>
<td>HDF5 Files</td>
<td>RGB Images</td>
<td>78.9 GB</td>
</tr>
<tr>
<td>14/11/2023 12:34</td>
<td>22/11/2023 13:55</td>
<td>24ZN8_FFAST</td>
<td>Phenomobile</td>
<td>HDF5 Files</td>
<td>RGB Images</td>
<td>50.0 GB</td>
</tr>
<tr>
<td>28/07/2023 12:00</td>
<td>22/11/2023 09:46</td>
<td>23TE43_cimson</td>
<td>Drone</td>
<td>RGB Images</td>
<td>26.0 GB</td>
<td></td>
</tr>
<tr>
<td>16/06/2023 09:23</td>
<td>22/11/2023 08:23</td>
<td>23ZM15_GIP</td>
<td>Phenomobile</td>
<td>HDF5 Files</td>
<td>RGB Images</td>
<td>Multispectral Images</td>
</tr>
<tr>
<td>20/07/2023 12:00</td>
<td>16/11/2023 14:03</td>
<td>23TE43_cimson</td>
<td>Drone</td>
<td>RGB Images</td>
<td>35.8 GB</td>
<td></td>
</tr>
<tr>
<td>14/11/2023 21:03</td>
<td>14/11/2023 20:05</td>
<td>24ZN8_FFAST</td>
<td>Phenomobile</td>
<td>HDF5 Files</td>
<td>RGB Images</td>
<td>Point Clouds</td>
</tr>
<tr>
<td>12/07/2023 12:00</td>
<td>09/11/2023 14:38</td>
<td>23TE43_cimson</td>
<td>Drone</td>
<td>RGB Images</td>
<td>20.1 GB</td>
<td></td>
</tr>
<tr>
<td>03/07/2023 12:00</td>
<td>09/11/2023 14:50</td>
<td>23TE43_cimson</td>
<td>Drone</td>
<td>RGB Images</td>
<td>32.5 GB</td>
<td></td>
</tr>
</tbody>
</table>
# Data Status

## Description
- **Experiment(s):** 3ZM15_GIP
- **Measuring session:** 16/06/2023 09:23
- **Loading date:** 22/11/2023 08:23
- **Size:** 277.9 GB
- **iRODS path:** /Francis/Gritze/Enz/home/4p/qualification/2780
- **Vector:** Phenomobile
- **PHIS:**

## Data

<table>
<thead>
<tr>
<th>Name</th>
<th>Version</th>
<th>State</th>
<th>PHIS</th>
<th>Size</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td>HDR Files</td>
<td>v1</td>
<td></td>
<td></td>
<td>63.4 GB</td>
<td></td>
</tr>
<tr>
<td>RGB Images</td>
<td>v1</td>
<td></td>
<td></td>
<td>40.5 GB</td>
<td></td>
</tr>
<tr>
<td>Multiplexed Images</td>
<td>v1</td>
<td></td>
<td></td>
<td>15.9 GB</td>
<td></td>
</tr>
<tr>
<td>Plant Clouds</td>
<td>v1</td>
<td></td>
<td></td>
<td>121.7 GB</td>
<td></td>
</tr>
<tr>
<td>Cropped RGB Images</td>
<td>v1</td>
<td></td>
<td></td>
<td>9.1 GB</td>
<td></td>
</tr>
<tr>
<td>VegBandy Seg RGB Images</td>
<td>v1</td>
<td></td>
<td></td>
<td>6.9 GB</td>
<td></td>
</tr>
<tr>
<td>GreenGal Seg RGB Images</td>
<td>v1</td>
<td></td>
<td></td>
<td>16.5 GB</td>
<td></td>
</tr>
<tr>
<td>Dac Fraction</td>
<td>v1</td>
<td></td>
<td></td>
<td>1.0 GB</td>
<td></td>
</tr>
<tr>
<td>Canopy Height</td>
<td>v1</td>
<td></td>
<td></td>
<td>194 KB</td>
<td></td>
</tr>
<tr>
<td>Canopy CoverFraction</td>
<td>v1</td>
<td></td>
<td></td>
<td>327 KB</td>
<td></td>
</tr>
<tr>
<td>Canopy GreenGal FVC</td>
<td>v1</td>
<td></td>
<td></td>
<td>419 KB</td>
<td></td>
</tr>
</tbody>
</table>

## Task history

<table>
<thead>
<tr>
<th>Macro</th>
<th>Label</th>
<th>Input data</th>
<th>Output data</th>
<th>Author</th>
<th>Start date</th>
<th>Duration</th>
<th>Status / Progress</th>
<th>Actions</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Phenom GreenGal FVC</td>
<td>RGB Images.v1</td>
<td>GreenGal Seg RGB Images.v1</td>
<td>Rémy Marandel</td>
<td>01/12/2023</td>
<td>02:33min</td>
<td>Complete</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>RGB Images.v1</td>
<td>Canopy GreenGal FVC.v1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Macro</td>
<td>Phenom CoverFraction</td>
<td>RGB Images.v1</td>
<td>VegBandy Seg RGB Images.v1</td>
<td>Rémy Marandel</td>
<td>01/12/2023</td>
<td>05:41min</td>
<td>Complete</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>RGB Images.v1</td>
<td>Canopy CoverFraction v1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Data processed with that pipeline
Manage versions
Display Processing chain
And each Module input/output
Module Ex: Export 4P data to PHIS

**Description**: Module d'export des données L2 vers PHIS

**Compatible species**: All

**Type**: Export module

**Author**: Antony Tong (antony.tong@agthesis-consult.com)

**Compatible vectors**: Module - export L2 PHIS (v1)

**Technical guide**: Module - export L2 PHIS (v1)

**Source repository**: git@opencms.inra.fr:dev-modules/phis-export-l2.pl

**Docker image**: phis-export-l2:1.0.0

**Status**: Published from 03/01/2023 10:37

**Integration status**: Published from 03/01/2023 10:39

**Inputs**

- **Prefix**: Current Installation
  - Type: string

- **-experiment**: Experiment Ids
  - Type: string

- **-json**: Configuration file
  - Type: file

- **-provenanceId**: Provenance URI
  - Type: string

- **-sensorId**: Sensor URI
  - Type: string

- **-macro**: Macro ID
  - Type: integer

- **-archiveId**: Archive URI
  - Type: string
## Catalog

<table>
<thead>
<tr>
<th>Name</th>
<th>Format</th>
<th>Size</th>
<th>Category</th>
<th>Vector</th>
<th>Organization</th>
<th>Status</th>
</tr>
</thead>
<tbody>
<tr>
<td>Plant segmentation model configuration</td>
<td>JSON file</td>
<td>3 KB</td>
<td>All</td>
<td>Setting</td>
<td>Phenomobile, Literal</td>
<td>Toulouse</td>
</tr>
<tr>
<td>Plant segmentation model weights</td>
<td>PTH file</td>
<td>40.7 MB</td>
<td>All</td>
<td>Deep Learning Mode</td>
<td>Phenomobile, Literal</td>
<td>Toulouse</td>
</tr>
<tr>
<td>Pheno app density model weights</td>
<td>PT file</td>
<td>175.1 MB</td>
<td>All</td>
<td>Deep Learning Mode</td>
<td>Phenomobile, Literal</td>
<td>Toulouse</td>
</tr>
<tr>
<td>GreenSen segmentation model</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>kmodel</td>
<td>CSV file</td>
<td>14 KB</td>
<td>All</td>
<td>Deep Learning Mode</td>
<td>Phenomobile</td>
<td>Toulouse</td>
</tr>
<tr>
<td>Eardensity model weights</td>
<td>PT file</td>
<td>175.1 MB</td>
<td>All</td>
<td>Deep Learning Mode</td>
<td>Phenomobile</td>
<td>Toulouse</td>
</tr>
<tr>
<td>White balance factors file RGB</td>
<td>CSV file</td>
<td>113 B</td>
<td>All</td>
<td>Calibration</td>
<td>Phenomobile</td>
<td>Toulouse</td>
</tr>
<tr>
<td>Soil height map</td>
<td>MAT file</td>
<td>1 KB</td>
<td>All</td>
<td>Measure</td>
<td>Drones</td>
<td>Toulouse</td>
</tr>
<tr>
<td>Transformation matrix 2 to 1</td>
<td>TXT file</td>
<td>406 B</td>
<td>All</td>
<td>Calibration</td>
<td>Phenomobile</td>
<td>Toulouse</td>
</tr>
<tr>
<td>Transformation matrix 3 to 1</td>
<td>TXT file</td>
<td>410 B</td>
<td>All</td>
<td>Calibration</td>
<td>Phenomobile</td>
<td>Toulouse</td>
</tr>
</tbody>
</table>

Showing 1 to 10 of 11 entries

1 2 Next
## List of available traits

<table>
<thead>
<tr>
<th>TRAIT</th>
<th>METHOD</th>
<th>SENSOR</th>
<th>VECTOR</th>
<th>REFERENCE</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Height Distribution</td>
<td></td>
<td>AR</td>
<td>Madec et al., 2017</td>
</tr>
<tr>
<td><strong>Vegetation Fraction (VF)</strong></td>
<td>DL segmentation</td>
<td></td>
<td>LI</td>
<td>Madec et al., 2022</td>
</tr>
<tr>
<td></td>
<td>Height threshold</td>
<td></td>
<td>AR</td>
<td>Lopez-Lozano et al., 2022</td>
</tr>
<tr>
<td><strong>Green Fraction (GF)</strong></td>
<td>VI Empirical</td>
<td>RG</td>
<td>Li</td>
<td>Jiang et al, 2018, Jay et al, 2019</td>
</tr>
<tr>
<td></td>
<td>ML &amp; DL segmentation</td>
<td></td>
<td>AR</td>
<td>Serouart et al., 2022, Madec et al., 2022</td>
</tr>
<tr>
<td></td>
<td>1D RTM inversion</td>
<td>RG</td>
<td>LI</td>
<td>Djamai et al, 2019, Camacho et al, 2021</td>
</tr>
<tr>
<td><strong>Green Area Index (GAI)</strong></td>
<td>VI Empirical</td>
<td>RG</td>
<td>AR</td>
<td>Jiang et al, 2018, Jay et al, 2019, Camacho, 2021</td>
</tr>
<tr>
<td></td>
<td>1D RTM inversion</td>
<td>RG</td>
<td>AR</td>
<td>Djamai et al, 2019, Jay et al, 2019, Camacho et al, 2021</td>
</tr>
<tr>
<td><strong>Plant Area Index (PAI)</strong></td>
<td>1D Turbid</td>
<td></td>
<td>AR</td>
<td>Lopez-Lozano et al., en prep</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>TRAIT</th>
<th>METHOD</th>
<th>SENSOR</th>
<th>VECTOR</th>
<th>REFERENCE</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Average Inclination Angle (AIA)</strong></td>
<td>1D RTM inversion</td>
<td></td>
<td>AR</td>
<td>Liu et al., 2017, Lopez-Lozano et al., en prep, Liu et al, 2019</td>
</tr>
<tr>
<td></td>
<td>3D RTM inversion</td>
<td></td>
<td>AR</td>
<td>Jiang et al., 2017</td>
</tr>
<tr>
<td><strong>Canopy Chlorophyll Content (CCC)</strong></td>
<td>1D RTM inversion</td>
<td></td>
<td>AR</td>
<td>Delloye et al, 2018</td>
</tr>
<tr>
<td></td>
<td>Vi Empirical</td>
<td></td>
<td>AR</td>
<td>Jay et al., 2019</td>
</tr>
<tr>
<td></td>
<td>3D Distribution of Leaf Area</td>
<td>1D Turbid</td>
<td></td>
<td>Liu et al., 2017, Lopez-Lozano et al., en prep</td>
</tr>
<tr>
<td><strong>Plant density</strong></td>
<td>DL</td>
<td></td>
<td>AR</td>
<td>Jin et al., 2017, Velumani et al, 2021</td>
</tr>
<tr>
<td><strong>Ear density</strong></td>
<td>DL @ reprod, stage</td>
<td></td>
<td>AR</td>
<td>Madec et al., 2019</td>
</tr>
<tr>
<td><strong>Leaf Chlorophyll Content</strong></td>
<td>1D RTM inversion</td>
<td></td>
<td>AR</td>
<td>Jiang et al, 2018</td>
</tr>
<tr>
<td></td>
<td>Vi ML, Empirical</td>
<td></td>
<td>AR</td>
<td>Jay et al., 2017, 2019, Jay et al, en prep</td>
</tr>
</tbody>
</table>
RGB Images/ Green Fraction

1 - Vegetation/Background Masks

2 - Green/Senescent vegetation

Serouart et al, 2022
Madec et al, 2023
RGB Images+ LiDAR / Ear Density

DL model from the global wheat head data challenge : David et al, 2021, 2023
Footprint of the RGB camera derived from camera height estimated from LidAR

RMSE = 46épis/m²

R² = 0.72

Densité mesurée (manuel)

Pre-FFAST PHENOV2
LiDAR / Height / Gap Fraction / Plant Area Density

Lopez-Lozano et al, in prep

Pre-FFAST project (2021) @ Diascope.
10 G x well-watered x November sowing x 2 replicates
THE TEAM

Platform Development
Software and functional architecture

France Grille
Cloud computing and storage

Platform Management
Algorithms
Processing pipelines

UMR EMMAH

INRAE

UMR MISTEA

Information System

Experimental Units
UE GCA
UE PHACC
UE DIASCOPE

• Specifications
• Test
• Use

THANKS FOR YOUR ATTENTION

EMPHASIS

EUROPEAN INFRASTRUCTURE FOR PLANT PHENOTYPING
Data enrichment: Events, Annotation and Documents

Definitions

- **Events**: can be processes, actions or facts that occur in (or even precede) scientific experiments and have an influence on the experiment. Events are identified and characterized. Main categories are controlled (irrigation, fertilizer, maintenance, installation) or uncontrolled (hail, frost, pests, breakdown).

  => Treatment, Trouble, Pest attack, Move, etc.

- **Annotation**: A short comment added to an entity (data, object, experiment, device, etc.).

  => Motivations: Comment, Highlight, Classify, Moderate, Link, Tag, etc.

- **Documents**: A document refers to a combination of a medium and information (the content), the latter recorded in a persistent manner. It has explanatory, descriptive or evidential value.

  => Report, Publication, Technical document, Experimental protocol, Dataset, etc.
Data enrichment: Events, Annotation and Documents

Definitions

• **Events:** can be processes, actions or facts that occur in (or even precede) scientific experiments and have an influence on the experiment. Events are identified and characterized. Main categories are controlled (irrigation, fertilizer, maintenance, installation) or uncontrolled (hail, frost, pests, breakdown).

  => Treatment, Trouble, Pest attack, Move, etc.

  => MIAPPE, Ontology of Experimental Events (OEEV) - [https://agroportal.lirmm.fr/ontologies/OEEV](https://agroportal.lirmm.fr/ontologies/OEEV)

• **Annotation:** A short comment added to an entity (data, object, experiment, device, etc.).

  => Motivations: Comment, Highlight, Classify, Moderate, Link, Tag, etc.

  => Web Annotation Ontology (OA) - [https://www.w3.org/ns/oa](https://www.w3.org/ns/oa)

• **Documents:** A document refers to a combination of a medium and information (the content), the latter recorded in a persistent manner. It has explanatory, descriptive or evidential value.

  => Report, Publication, Technical document, Experimental protocol, Dataset, etc.

  => Dublin Core - [https://www.dublincore.org/specifications/dublin-core/dcmi-terms/](https://www.dublincore.org/specifications/dublin-core/dcmi-terms/)
Data enrichment: Events, Annotation and Documents

Definitions

• **Events**: can be processes, actions or facts that occur in (or even precede) scientific experiments and have an influence on the experiment. Events are **identified and characterized**. Main categories are controlled (irrigation, fertilizer, maintenance, installation) or uncontrolled (hail, frost, pests, breakdown).

  => Treatment, Trouble, Pest attack, Move, etc.

  => MIAPPE, Ontology of Experimental Events (OEEV) - https://agroportal.lirmm.fr/ontologies/OEEV

• **Annotation**: A short comment added to an entity (data, object, experiment, device, etc.).

  => Motivations: Comment, Highlight, Classify, Moderate, Link, Tag, etc.

  => Web Annotation Ontology (OA) - https://www.w3.org/ns/oa

• **Documents**: A document refers to a combination of a medium and information (the content), the latter recorded in a persistent manner. It has explanatory, descriptive or evidential value.

  => Report, Presentation, Technical document, Experimental protocol, Dataset, etc.

  => Dublin Core - https://www.dublincore.org/specifications/dublin-core/dcmi-terms/

Add more value with semantic metadata!

Ontologies, thesaurus, etc.
Data enrichment: Events, Annotation and Documents

Illustration

- **Event356 Storm**
  - **Plot97**
  - **anno480**
    - oa:hasTarget
    - dcterms:creator
    - dcterms:created: "2018-07-06"
    - "Plots lodged after the storm"

- **Event849 Pot Fall**
  - **Plant795**
  - **anno480**
    - oa:hasTarget
    - dcterms:creator
    - dcterms:created: "2018-05-26"
    - "Fallen plant during imaging"
Events

Main Event Types

- Displacement
  - Move from
  - Move to

- Trouble
  - Breakdown
  - Incident
  - Dysfunction
  - Pest attack
  - Pot fall

- Facility Management
  - Calibration
  - Restart
  - Start
  - Stop

- Scientific Object Management
  - Irrigation
  - Sowing
  - Harvesting
  - Treatment
Events - MIAPPE important Section

Event

- Discrete, dated or time-stamped occurrence
- Natural (e.g. rain, pathogen attack)
- Cultural practice (e.g. sowing, irrigation)
- Applied to the whole study or by observation unit
- This is not the factor, but additional information.
- Event can be used to achieve a factor.

Metadata: name, description and time/date

Example: POPLAR [2]

- the field establishment date, 2003
- the orchard was subjected to 15mm of rain on March 15, 2012 (fiction)

<table>
<thead>
<tr>
<th>Study</th>
<th>Event</th>
<th>Name</th>
<th>Description</th>
<th>Date</th>
</tr>
</thead>
<tbody>
<tr>
<td>Monclus et al., 2012</td>
<td>Rain</td>
<td></td>
<td>15mm of rain on the orchard</td>
<td>2012-03-15</td>
</tr>
</tbody>
</table>
Events

Event Types & Properties

- Event ID
- Event Type
- Event Description
- Event Date / Period
- Target
Events

Event Types & Properties: Move

- Event ID
- Event Type
- Event Description
- Event Date / Period
- Target
- From
- To
- Coordinates
- X, Y, Z Position
- Textual position
Events

Event Types & Properties: Treatment

- Event ID
- Event Type
- Event Description
- Event Date / Period
- Target
- Product
- Quantity
- Pest
- Equipment used
- Agent

Plot57

Event596
Pest Attack

“2023-06-01”
‘slug damage’

Event Treatment

Anti-slug Ferramol

“2023-06-02”

Pulverizer2

Agent5

Agricultural Protocol
Slug control product Ferramol

2023-06-01

2023-06-02
Events

Describe a scientific object life cycle

Definition and declaration of Pots  2017-01-12
Sowing  2017-02-10
Definition and declaration of Plants  2017-04-12
Move  2017-06-10
Harvesting  2017-09-28
Data Exploration: environmental and phenotypic data

Visualising and linking environmental and phenotypic data
Data Integration Architecture

Unified View Model

Query as if you had a single source!

Unified Schema

Data virtual integration

SPARQL EndPoint
RDF connector
NoSQL connector

EMPHASIS
EUROPEAN INFRASTRUCTURE FOR PLANT PHENOTYPING

37
Use Case: Data analysis and integration

How FAIR data may help in managing and analyzing data in HTPP experiments? - Llorenç Cabrera Bosquet
Data analysis and integration

How FAIR data may help in managing and analyzing data in HTPP experiments?

First -> Some prerequisites

- Data are organised and managed following FAIR principles
- All data includes phenotypic, environmental and metadata measurements
- Ideally, data is managed in ‘real time‘ (allows decision-taking)
Exemple 1: Piloting and monitoring water deficit in thousands of plants

We need to continuously and individually monitor and manage water scenarios in most experiments.

PHENOARCH platform Montpellier, INRAE

Well-watered plants

Imposition of water deficit (ca. -4 bar)
How soil water content (SWC) and SWP is measured?

1. Plant are weighed individually on a daily basis (or more)

2. Soil water content (g H₂O per g of dry soil) is deduced using a formula

3. Soil water potential (bar) is calculated using a soil retention curve (Ψ vs Φ)

\[
\text{SWC} = \text{SW} \times a + b
\]

Soil weight

Soil water content (g g⁻¹)

Soil water potential (bar)

van Genuchten adjustment (1980)

\[ n = 1.1611, \alpha = 63.355 \]
This is not so easy...

1. Platform weighs the whole ‘system’, not only soil weight
This is not so easy...

1. Platform weighs the whole ‘system’, not only soil weight

WE NEED

Proper identification of individual components

URI of plant
<m3p:arch/2017/c17000118>

URI of pot:
<m3p:arch/2013/po13001542>

URI of cart:
<m3p:arch/2013/ct1300123>
This is not so easy...

1. Platform weighs the whole ‘system’, not only soil weight

**WE NEED**

Proper identification of individual components

- URI of plant: `<m3p:arch/2017/c17000118>`
- URI of pot: `<m3p:arch/2013/pc13001542>`
- URI of cart: `<m3p:arch/2013/ct1300123>`

Weight of individual components + metadata

- Weight of pots, carts, soil, etc.
- Soil used and its characteristics
- Initial soil water content at potting
- Events that may have affect the weight (e.g. fertilisation, adding of sticks, etc.)
This is not so easy...

2. Plants grow over time

This may have a HUGE impact on soil water content estimation if plant biomass is not taken into account
This is not so easy...

2. Plants grow over time

This may have a HUGE impact on soil water content estimation if not taken into account

**WE NEED**

Identification of relevant information

URI of plant
<m3p:arch/2017/c17000118>

URI of cabin:
<m3p:arch/2018/ac180015>

URI of camera:
<m3p:arch/2018/ac180019>

URI of image:
<m3p:arch/2017/ic17002295855>
This is not so easy...

2. Plants grow over time

Identification of relevant information

URI of plant
<m3p:arch/2017/c17000118>

URI of cabin:
<m3p:arch/2018/ac180015>

URI of camera:
<m3p:arch/2018/ac180019>

URI of image:
<m3p:arch/2017/ic17002295855>

Plant biomass estimation over time

Requires real time monitoring and analysis of data
This is not so easy...

2. Plants grow over time

WE NEED

Identification of relevant information

URI of plant
<m3p:arch/2017/c17000118>

URI of cabin:
<m3p:arch/2018/ac180015>

URI of camera:
<m3p:arch/2018/ac180019>

URI of image:
<m3p:arch/2017/ic1700295855>

Plant biomass estimation over time

Outliers

Requires real time monitoring and analysis of data
Finally SWC and SWP are calculated

\[
\text{SWC} = \text{Weight of ‘whole system’} - \text{Plant biomass} - \text{Weight of Pot} - \text{Weight of Cart} - \frac{b}{a}
\]

\[
\text{SWP} = \left(\frac{\text{SWC}}{2}\right)^{-\frac{n}{(n-1)}} - 1^{\frac{1}{n}} \left(-\alpha\right) / 10
\]
Envirotyping: Spatial and temporal variability and consequences for phenotyping

- Climate information is an essential **co-variate for monitoring growth** as drives most growth processes.
- Most platforms display a large temporal and spatial variability of environmental conditions

![Graphs showing PAR distribution](image)

**Granier et al 2006**

**4PMI Dijon**
Julien Martinet,
Christophe Salon

**RootPhair,**
UCL
Xavier Draye

There are some consequences for phenotyping...

**IGNORING IT** => depreciates the value of phenotypic data.

**MEASURING IT** enables

=> The **combination of data** from different experiments, different platforms or from field and platforms.

=> Taking advantage of (undesired) climate variability to determine **environmental response curves** during an experiment.
Exemple 2: Environmental quantification

**AIM**: Calculate thermal time from sowing for each plant in an experiment

\[ \text{ThermalTime} = \sum_{i=1}^{n} (T_{\text{air}} - T_{b}) \]

Apparently easy but…

- Plants sown and harvested in batches at different dates
- Plants belonging to different crops
- Plants located in different positions of the greenhouse
- Sensors at different positions displaying significant variation in temperature
Exemple 2: Environmental quantification

AIM: Calculate thermal time from sowing for each plant in an experiment

\[ \text{ThermalTime} = \sum_{i}^{n} (T_{air} - T_{b}) \]

WE NEED

- Plants sown in batches at different dates => Sowing and harvest dates for each plant (events)
- Plants belonging to different crops => \( T_{b} \) specific for each crop (ex. 8°C for maize, 0°C for wheat)
- Plants located in different positions of the greenhouse => tracking of plants’ position
- Sensors displaying significant variation in temperature => Sensor position, variable measured and units
Exemple 2: Environmental quantification

Taking into account all this annotated date allows to calculate thermal time for each plant in an experiment.
Statistical methods
Incorporation of statistical methods for platform management, data 'cleaning,' and analysis

1. **Detection of outlier plants** (bad germination, disease, seed error) 
   Alvarez Prado et al. 2019

2. **Design Generator**: User friendly online service aimed at construction and generation of experimental designs  
   [https://eppn2020design.com/](https://eppn2020design.com/)

3. **statgenHTP R package** (outlier detection, correction of spatial trends, parameter extraction from time series)
   [https://cran.r-project.org/web/packages/statgenHTP/](https://cran.r-project.org/web/packages/statgenHTP/)
   [https://biometris.github.io/statgenHTP/articles/vignettesSite/Intro_HTP.html](https://biometris.github.io/statgenHTP/articles/vignettesSite/Intro_HTP.html)
Spatio-temporal analysis of HTTP data

1. Design features and spatial trends correction per time point.

2. Longitudinal modelling of the spatially corrected data, thereby taking advantage of shared longitudinal features between genotypes and plants within genotypes.

3. Extracting time-independent attributes to characterise genotypes.

Code available
Geospatial Events

Event with an Area as a Target

- Event ID
- Event Type
- Event Description
- Event Date / Period

- Target = Area
  - Type
  - Name
  - Description
  - Geometry
Geospatial Events

Event with an Area as a Target

• Event ID
• Event Type
• Event Description
• Event Date / Period
• Target = Area
  • Type
  • Name
  • Description
  • Geometry

Events and also...

• Any observed object
• Any device
• Any facility
• Adresse (site, institution, ...)
• ...

EUROPEAN INFRASTRUCTURE FOR PLANT PHENOTYPING
IS vs GIS: How to deal with spatial data?

**Vector**
- Import
- Export
- Generation
- Update
- Attribute table

**Raster**
- Import
- Export
- Generation
- Processing

**Georeferencing**
- Projection sys
- Set up
- Up date
- Change
- Search & Geolocalisation

**Maps**
- Generation
- Symbology
- Atlas
- Layout

**Analysis**
- Filter
- Query
- Measurements
- Geostatistics
IS vs GIS: How to deal with spatial data?

Do we need a GIS to manage properly and FAIRly Geospatial phenotyping data?
What we need as minimum

• GIS: What are the limits?
  • File sizes, Transformation & Projection ⇒ access limits
  • DBMS: mostly relational
    • Limited performance and capacity in Big Data applications and MetaData management
    • Limited flexibility to handle complex data types and relationships
What we need as minimum

• GIS: What are the limits?
  • File sizes, Transformation & Projection ⇒ access limits
  • DBMS: mostly relational
    • Limited performance and capacity in Big Data applications and MetaData management
    • Limited flexibility to handle complex data types and relationships

Make a high-performance I.S. interoperable with a GIS
What we need as minimum

• **GIS**: What are the limits?

  • File sizes, Transformation & Projection ⇒ access limits
  • DBMS: mostly relational
    • Limited performance and capacity in Big Data applications and MetaData management
    • Limited flexibility to handle complex data types and relationships

---

**Make a high-performance I.S. interoperable with a GIS**

• Query & Analyse
• Exporting different formats (shapefiles, geojson, CSV,...)
• Spatial functions and indexation
• DB: PostGreSQL and extension PostGIS, Oracle Spatial, MongoDB, MySQL Spatial
What we need as minimum

- Vector
  - Import
  - Export
  - Generation
  - Update
  - Attribute table

- Raster
  - Import
  - Export
  - Generation
  - Processing

- Georeferencing Projection sys
  - Set up
  - Update
  - Change
  - Search & Geolocalisation

- Maps
  - Generation
  - Symbology
  - Atlas
  - Layout

- Analysis
  - Filter
  - Query
  - Measurements
  - Geostatistics
How to integrate spatial data (PHIS example)
How to integrate spatial data (PHIS example)

• Implementation examples:

Spatial data storage in **MongoDB** using **SOSA** ontology

### MongoDB:
- Document oriented DBMS (Object = document, organized in collections)
  - NoSQL
- Data format: BSON (binary JSON)
- Spatial data format: GeoJson (WGS84)

### SOSA Ontology (Semantic Sensor Network Ontology):
- Formal but lightweight general specification
- Modeling the interaction between entities
  - Sensor
  - Observation
  - Sample
  - Actuator
- Use the observation collection of SOSA:
  - 1 spatial data of 1 object = 1 observation
  - 1 list of spatial data of 1 object = 1 observation collection
How to integrate spatial data (PHIS example)

GeoJSON document = sosa: observation

"Location" collection= sosa:observationCollection
How to integrate spatial data (PHIS example)

- **Spatial Data format**
  - Projection system: mostly WGS84 (World Geodetic System 1984 - EPSG:4326) : longitude/latitude
  - GeoJson, WKT, KML, ....
  - Visualization libraries: OpenLayers, MapBox, LeafLet

```
GeoJSON

```

```
<?xml version="1.0" encoding="UTF-8"?>
<kml xmlns="http://www.opengis.net/kml/2.2">
<Document>
  <Placemark>
    <name>New York City</name>
    <description>New York City</description>
    <Point>
      <coordinates>-74.006393,40.714172</coordinates>
    </Point>
  </Placemark>
</Document>
</kml>

KML

```

```
POINT(6 10)
LINESTRING(3 4,10 50,20 25)
POLYGON((1 1,5 1,5 5,1 5,1 1))
```

WKT
Geospatial targets (Events: PHIS)

Example: Irrigation location - Draw the area
Geospatial targets (Events: PHIS)

Example: Irrigation location - Define the area
Geospatial targets (Events: PHIS)

Example: Irrigation location - Visualize the area
Geospatial targets (Objects: PHIS)
PHIS to QGIS (objects, devices, events)
And then?

- I.S. managing all data including spatial data
  - Import and export
  - Query
  - Visualization
  - Interoperable with GIS

⇒ Meta analysis & Geostatistic

⇒ Historical data as well as data for the current year
Geostatistics & Spatial tools: Zoning & beyond

Mostly used for yield data ... and other spatial data!

- More and more agronomic variables
- Possibility of adaptive management based on intra-field variability
  - How and where implement an experiment?
  - How to filter and clean spatial data?
  - How to validate the relevance of the outlier detection and filtering?
  - How to integrate expertise (agricultures, advisors, ...)?
  - How to simulate a spatial data set based on historical and current data?
  - How to generate recommendation maps (Decision-Support Tools)
Geostatistics & Spatial tools: Outlier detection & filtering

Mostly used for yield data ... and other spatial data! (C. Leroux & H. Jones)
Outlier detection, filtering & zoning (fertilization)

Mostly used for yield data ... and other spatial data! (C. Leroux & H. Jones)

- Outlier deleting
- Reassessment of yield characteristics
- Operational zoning for yield data (clustering)
- Fertilization application
Geostatistics & Spatial tools: Yield map simulation

Mostly used for yield data ... and other spatial data! (C. Leroux & H. Jones)
Geostatistics & Spatial tools

Zoning: Historical phenotyping data for experiment implementation

- Maximizing:
  - Inter zones heterogeneity
  - Intra zone homogeneity

- Shape constraints

Variable: Yield, Soil parameters, Biomass, ...
Geostatistics & Spatial tools: Zoning

Zoning: Historical phenotyping data for experiment implementation

Value: Yield (N-1)

(C. Leroux & H. Jones)
Geostatistics & Spatial tools: Zoning

Zoning: Historical phenotyping data for experiment implementation

Value: Yield (N-1)

(C. Leroux & H. Jones)
Geostatistics & Spatial tools

Vineyard yield forecasting based on machine learning methods (H. Jones et al.,)

Data (historical & spatio-temporeal)
- Yield sensors
- Crop sensors (VIs)
- Soil sensors
- Pruning weight

Methode Random Forest

Machine learning algorithm
- Supervised learning (need for examples for inputs/output to learn)
- Possible to use on
  • Classification (predict a class)
  • Regression (predict a value)

What & When are the best data for yield reporting?
**Geostatistics & Spatial tools**

**Vineyard yield forecasting based on machine learning methods** (H. Jones et al.,)

What is to be predicted?

<table>
<thead>
<tr>
<th>Variable</th>
<th>2018</th>
<th>2019</th>
<th>2020</th>
<th>2021</th>
</tr>
</thead>
<tbody>
<tr>
<td>Pruning mass (from hand sampling post season)</td>
<td><img src="image" alt="image" /></td>
<td><img src="image" alt="image" /></td>
<td><img src="image" alt="image" /></td>
<td><img src="image" alt="image" /></td>
</tr>
<tr>
<td>Yield (from AgLeader yield monitor)</td>
<td><img src="image" alt="image" /></td>
<td><img src="image" alt="image" /></td>
<td><img src="image" alt="image" /></td>
<td><img src="image" alt="image" /></td>
</tr>
</tbody>
</table>
Geostatistics & Spatial tools

Vineyard yield forecasting based on machine learning methods (H. Jones et al.,)

Some results...

- Prediction on yield (2019)
  - Mean absolute error: 0.127
  - Mean absolute percentage error: 0.0187
  - explained_variance: 0.608
  - range y_test: [5.83 7.37]
  - feature importance:
    - SR_310519 0.079783
    - DfVI_170619 0.058389
    - MSR_170619 0.053485

- Prediction on Prune Weight (2020)
  - Mean absolute error: 0.887
  - Mean absolute percentage error: 0.174
  - explained_variance: 0.613
  - range y_test: [1.5 11.125]
  - feature importance:
    - PW_LbsperPanel_2019 0.281330
    - CropLoad2019_AgLeader 0.27707
Conclusion & Perspectives

- Think FAIR
- Structure Data
- Use Standards
- Think of MetaData
- Use Ontology
- Use semantic
- Use appropriate DBMS
- Use appropriate storage systems
- Make good DMP
- GDPR compliant
Conclusion & Perspectives

- Think FAIR
- Structure Data
- Use Standards
- Think of MetaData
- Use Ontology
- Use semantic
- Use appropriate DBMS
- Use appropriate storage systems
- Make good DMP
- GDPR compliant

Seems infeasible? Ask for help!

● Information systems
● Ontology
● Pipelines for data transmission & processing
● Mapping with standards
● DMP & GDPR setting
● Spatio-temporal data for advanced analysis (variables)
Conclusion & Perspectives

- Start changing habits and adopt appropriate methods & tools
- Trainings, workshops & Hackathons
  - PHENET-EMPHASIS Joint training event in 2024
    - Training materials and survey to set the dates and content
  - BrAPI Hackathon in 2024 - https://brapi.org/events/hackathon
  - BioHackathon in 2024 - https://biohackathon-europe.org/
  - PhenoHarmonIS - Montpellier May 2024
Conclusion & Perspectives

- Start changing habits and adopt appropriate methods & tools
- Trainings, workshops & Hackathons
  - PHENET-EMPHASIS Joint training event in 2024
    - Training materials and survey to set the dates and content
  - BrAPI Hackathon in 2024 - [https://brapi.org/events/hackathon](https://brapi.org/events/hackathon)
  - BioHackathon in 2024 - [https://biohackathon-europe.org/](https://biohackathon-europe.org/)
  - PhenoHarmonIS - Montpellier May 2024

Contact us:
- opensilex@gmail.com
- isabelle.alic@inrae.fr
- farzaneh.kazemipour@inrae.fr
- pascal.neveu@inrae.fr
- cyril.pommier@inrae.fr
Special thanks to: Silvana Moscatelli, François Tardieu, Alexia Chiavarino, Sarra Abidri, Anne Tireau, Hazaël Jones...

https://www.phenome-emphasis.fr/

OpenSILEX Team - http://opensilex.org/
EMPHASIS is an ESFRI-listed project.

EMPHASIS-PREP is funded by the European Union (Grant Agreement: 739514).