



Gestion et Intégration de Données de Phénotypage Haut Débit et Web Sémantique



Pascal Neveu
UMR Mistea INRA Montpellier
Département MIA

What is it?



Big Data Challenge

More and more Data!

- **Storage capacity, Network flow, etc.**
1 Gigabyte: \$400K in 1980, \$10K in 1990, \$1K in 1995, \$10 in 2000, \$0.01 in 2017
- **Heterogeneous devices, simulations, etc.**
- **Internet sources** (*Open, collaborative, participative,...*)

Make data valuable!

- **Decision support**
- **Knowledge discovery**
- **New services**

- *Population treatment → individualized treatment*
- *When data did not quite match what we expect!*
- *Which theories/models are consistent and which ones are not!*
- ...

Big Data in Agronomy

V characteristics

- **Volume:** massive data and **exponential growing size**
→ *hard to store, manage and analyze*
- **Variety, Vocabulary and Complexity:** different sources, scales, disciplines different semantics, schemas and formats etc.
→ hard to understand, combine, integrate
- **Velocity:** speed of data generation
→ have to be processed on line
- **Veracity, Validity, Variability,** Vulnerability, Volatility, Visibility, Visualisation, Vagueness, etc.
- **Value**

FAIR DATA

Findable, Accessible, Interoperable, Reusable

Findable: PID, rich metadata and indexed in portals

Accessible: open and standardized protocols (internet protocols), authentication* (if not open)

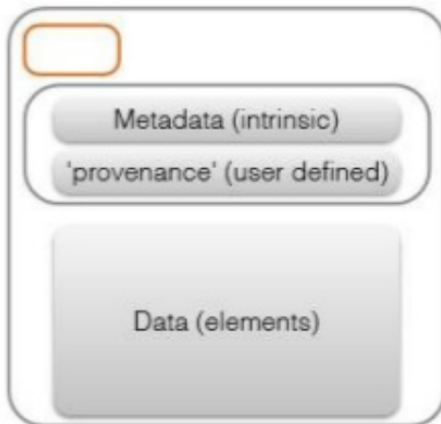
Interoperable: shared standardized formats and vocabularies (technology, syntax, semantic)

Reusable: provenance, domain relevant metadata for understanding

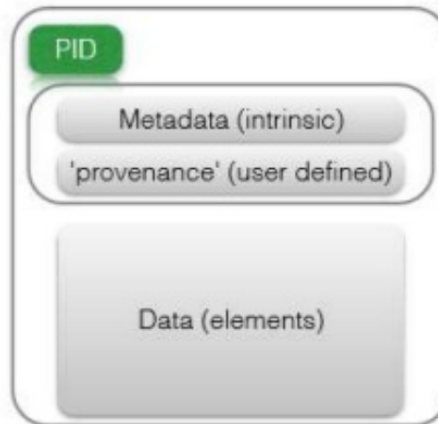
FAIR DATA

Findable, Accessible, Interoperable, Reusable

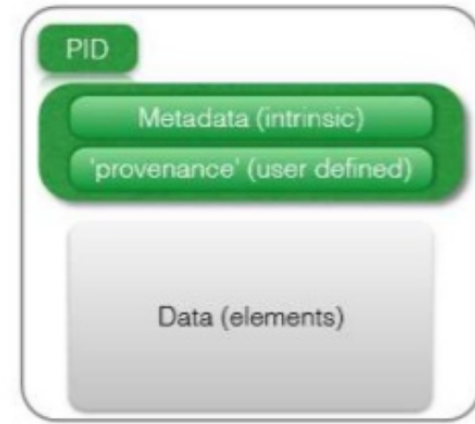
Totally UNFAIR



Findable
Usable for Humans



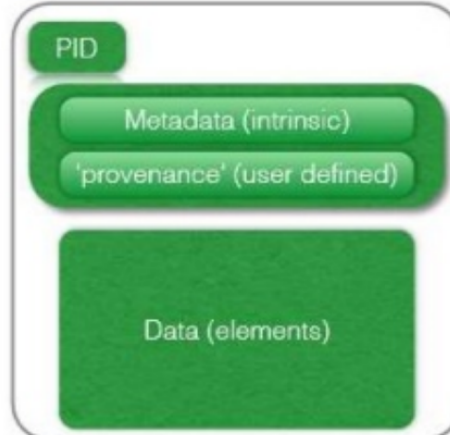
FAIR metadata



FAIR data-
restricted access



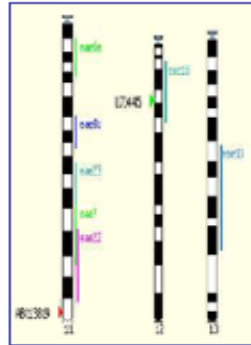
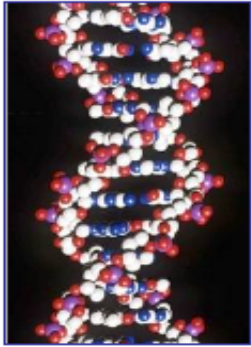
FAIR data-
Open Access



FAIR data-
Open Access/Functionally Linked



High Throughput Plant Phenotyping: searching for the most adapted genotypes



High frequency observations
of trait dynamics
for big set of Phenotypes

Many Plant **Genotypes**

Interactions



Various **Environments**



High Throughput Plant Phenotyping:

Decision support

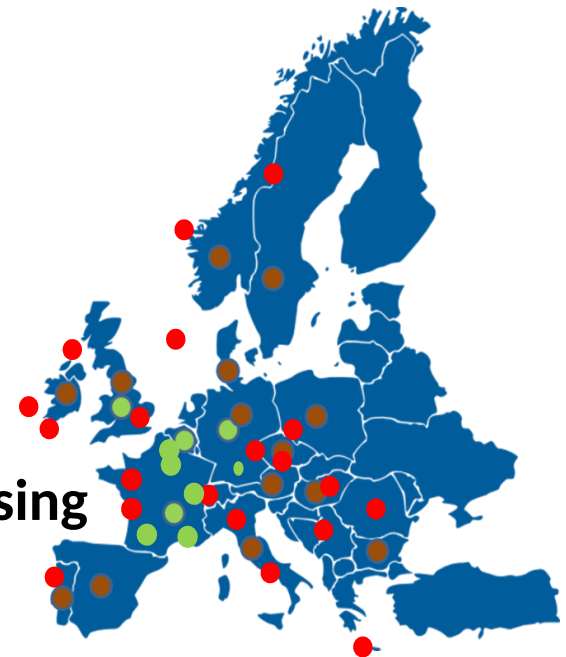
- Links genomics with plant ecophysiology and agronomy
- Phenotype-driven gene function discovery

Searching for the most adapted species/varieties for field challenges

- Food security
 - Climate Change adaptation
 - AgroEcology
 - Reduce inputs / natural resource preservation
 - **Safe and healthy food**
- Take into account food transformation and consumer

Emphasis e-infrastructure

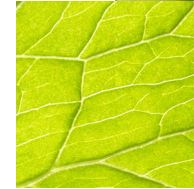
- Deals with several Petabytes of distributed data
- Makes FAIR data
- Based on Open technologies and standard (MIAPPE, BrAPI, etc)
- Standardized Identification
- Standardized Semantic
- Provenance and reproducibility data processing



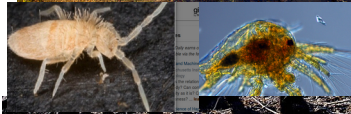
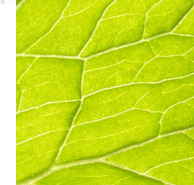
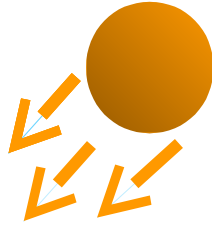
High Throughput Phenotypage data



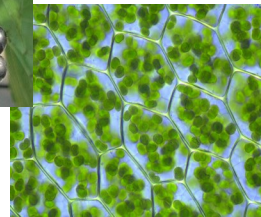
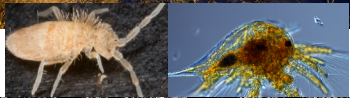
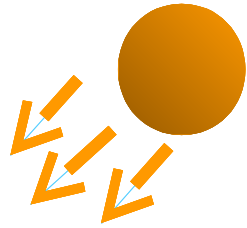
Different Scales



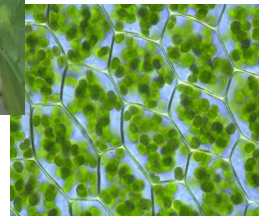
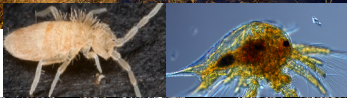
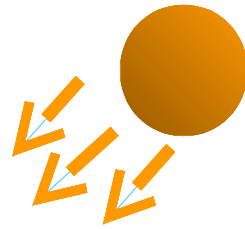
Interactions



Different Species



Different Stages and Transformations



From various contexts

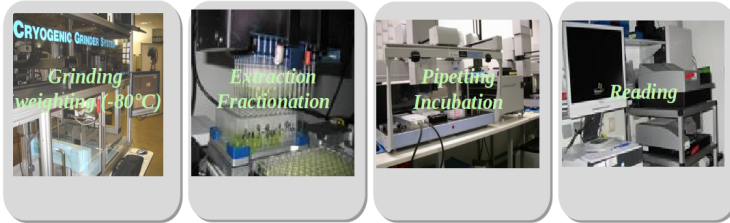
« omics » Platforms

Various data complex types

Genomics

Composition and the structure of biopolymers

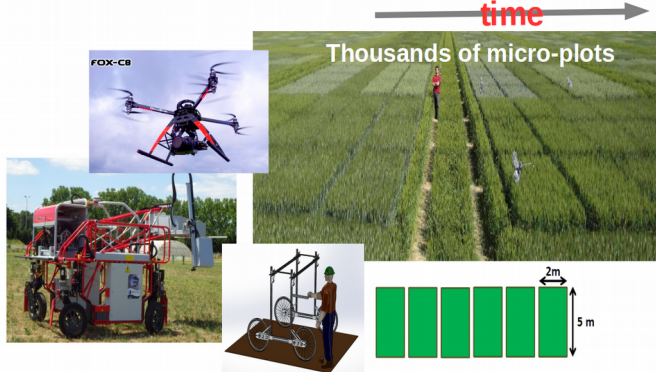
Quantification of metabolites and enzyme activities



Field Platforms

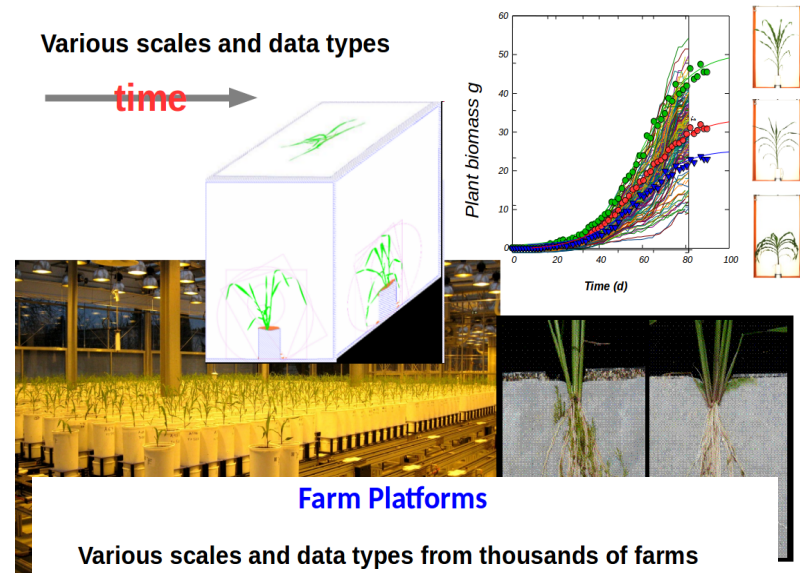
Various scales and data types

- Cell, organ, plant, population
- Images, hyperspectral, spectral, sensors, human readings...

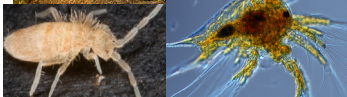
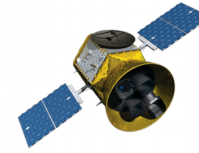
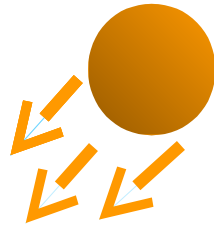


Green house Platforms

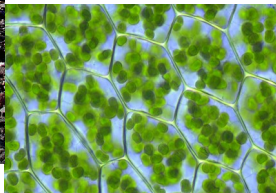
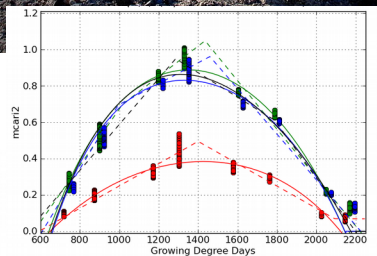
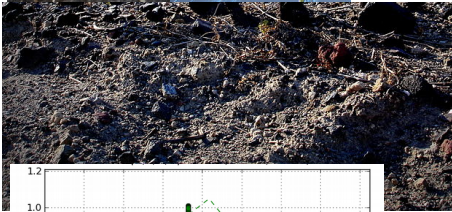
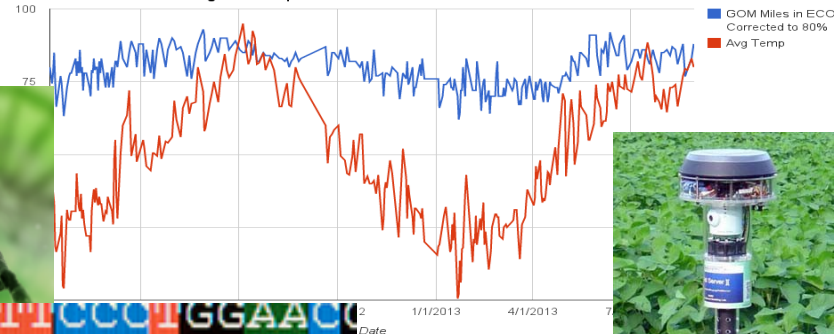
Various scales and data types



Heterogenous Data Sources



GOM Miles at 80% Charge and Temperature Data



How to structure data?



PHIS: An Information System for High Throughput Phenotyping

- ➔ Designed and developed for **smart** (Big) Data management of various **Phenotyping platforms** (greenhouse, field, farm)
- Management of huge, complex and heterogeneous data (millions of images, sensor data, etc)
- Implement good practices of data management
 - ✓ Make FAIR data
 - ✓ Foster collaborations (Open and Flexible)
 - ✓ Ability to understand and reproduce data processing
 - ✓ Ability to enforce DMP and Open Data



PHIS approaches

Ontology Driven

Scientific objects (plant, plant organ, plot, etc.) are:

- Identified by **URI** standardized, unambiguous, shared, etc
- Structured and linked with a controlled semantic (Ontology)
- **Events** (management, faults, meteo, etc) are semantically associated to these entities

A Specific platform **application Ontology (OWL*)** formalizes **Objects and Events (RDF*)** and allows to link **reference ontologies (SKOS*)**

Measurements, Documents, Observations, Metadata are associated with these Objects and Events

* Semantic Web languages

PHIS: Names and Identification

**Resource identification: standardized, unambiguous
Persistent?**

URI of plant :

`mp3:arch/2014/pl/000000012`

URI of pot :

`mp3:arch/2001/pt/000001542`

URI of cabin :

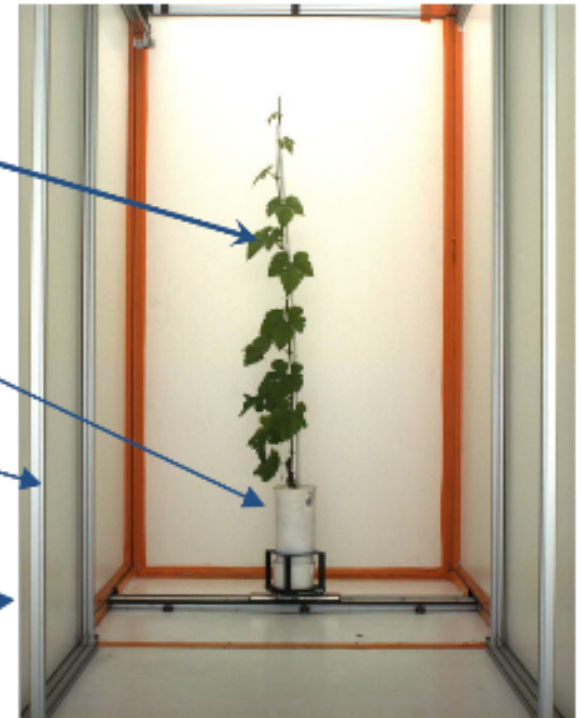
`mp3:arch/2010/ca/cabine2`

URI of camera :

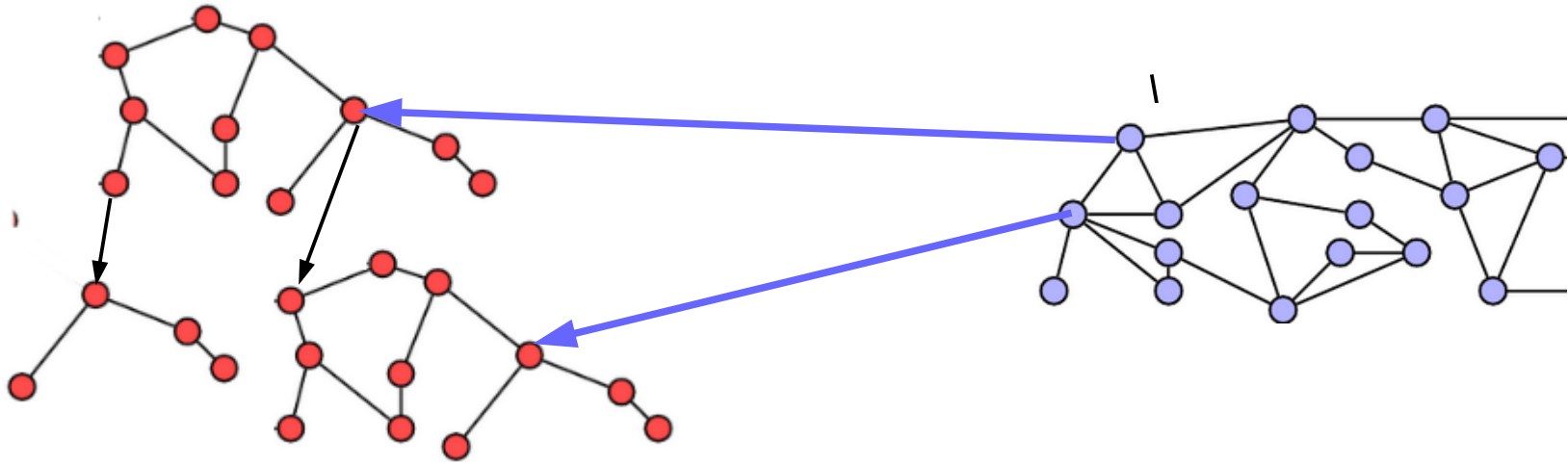
`mp3:arch/2011/ss/00003312`

URI of image :

`mp3:arch/2015/im/000000564`



Semantic Ressources



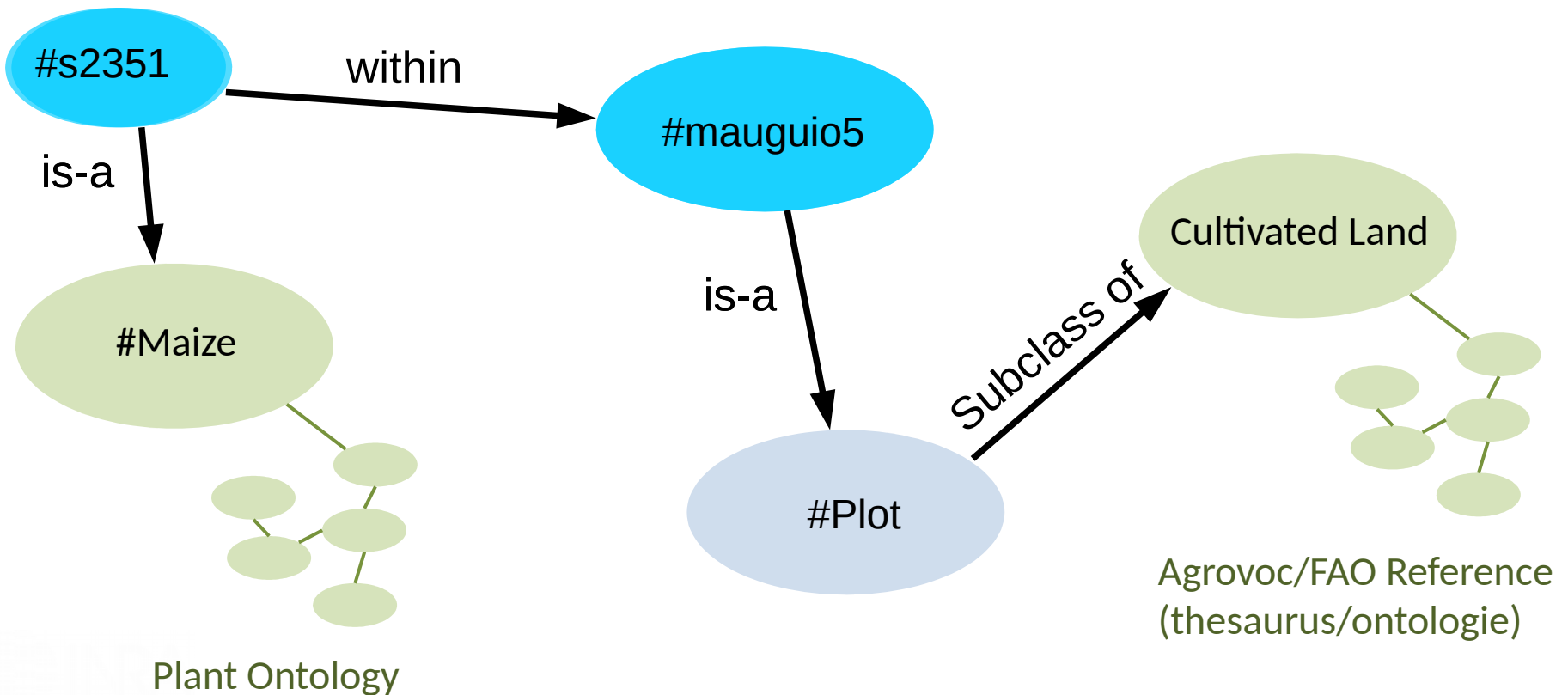
Reference ontologies

- Crop Ontology, Plant Ontology, TO, Agrovoc, EO

Application ontologies Web Annotation Ontology

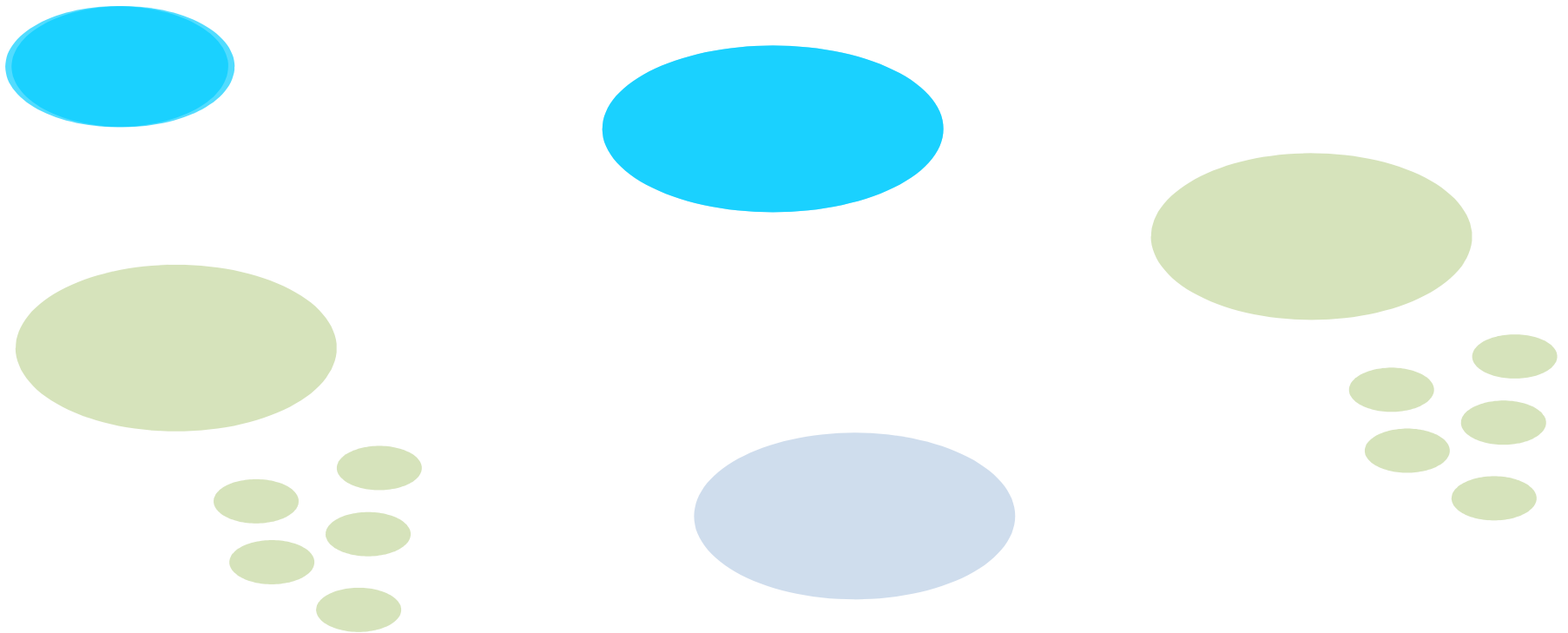
Semantics for the Structuring of Data

Metadata / Ontologies provide the meaning of the data
→ link each data element to a **controlled, shared vocabulary and *Machine readable***

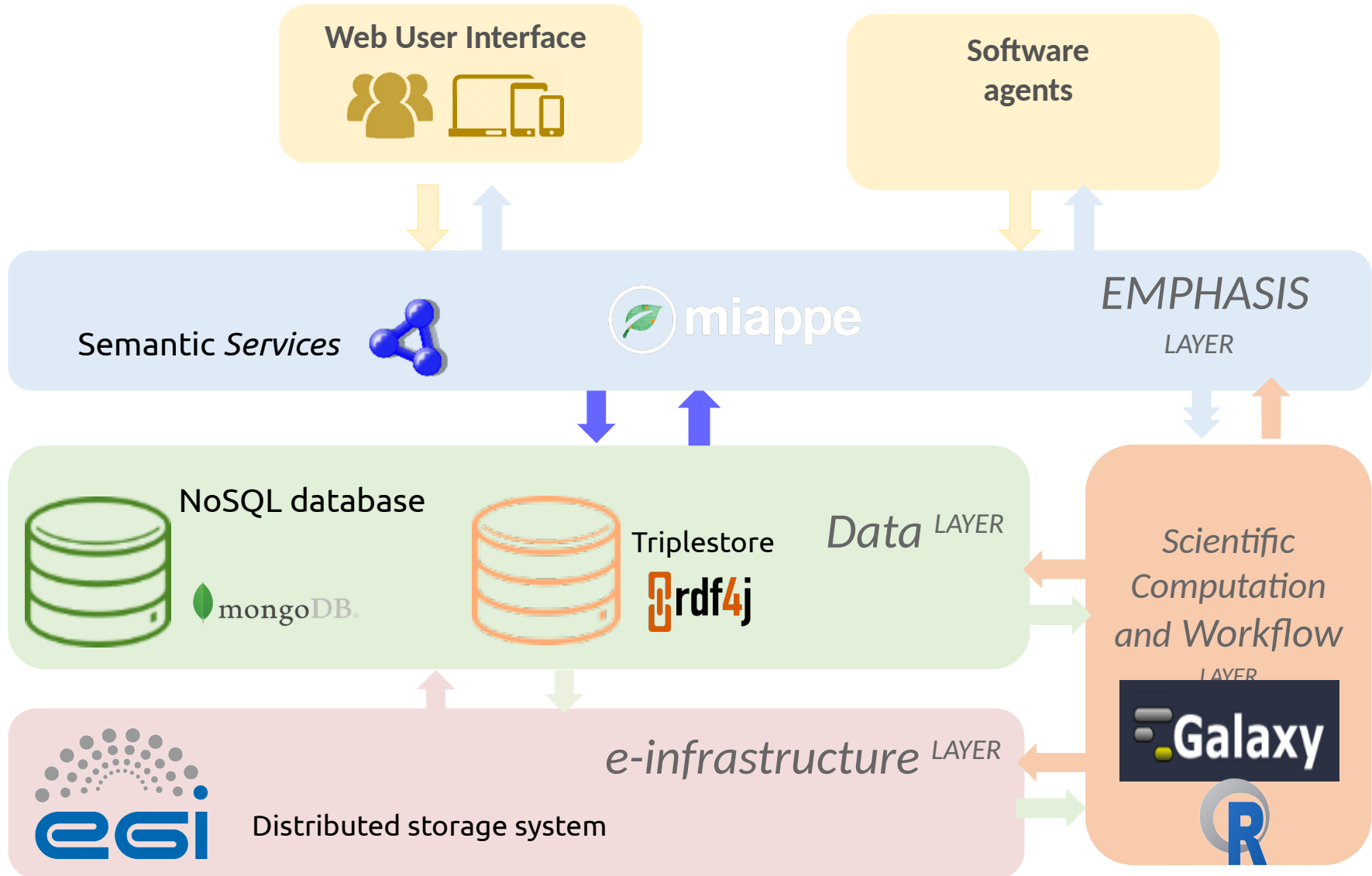


Semantics for the Structuring of Data

Metadata / Ontologies provide the meaning of the data



PHIS Architecture



Make easier Cloud computing

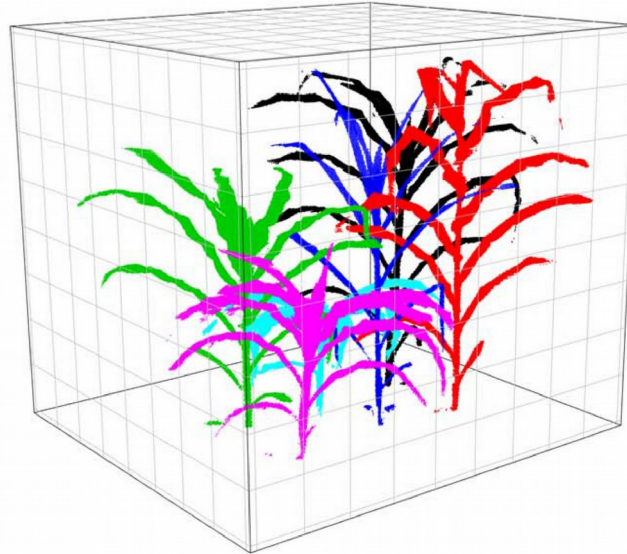
- **IRODS / ONeData**
(distributed storage of image datasets)
- **France Grille, EGI**
- **INRA Data Center**



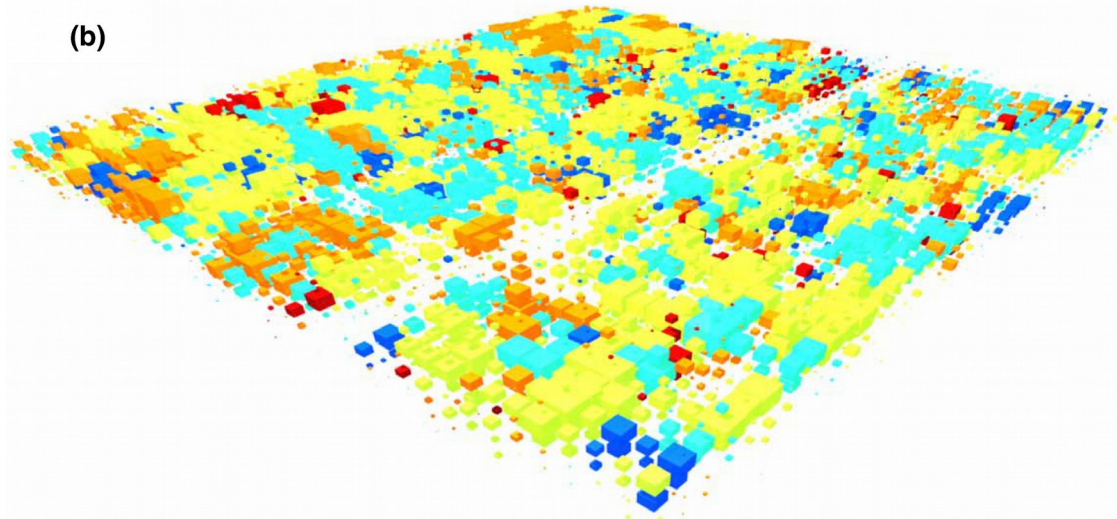
Knowledge Discovery Illustration

PHIS provides contextualisation: intercepted light value

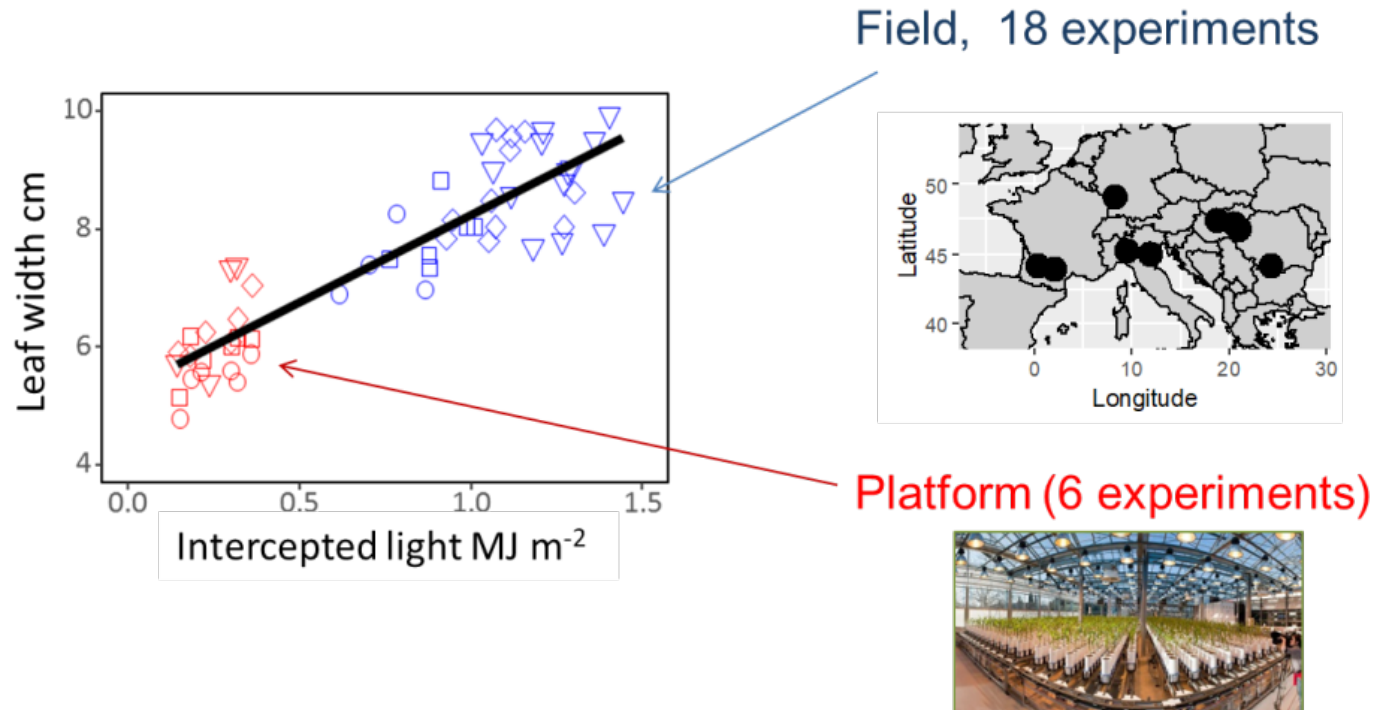
(a)



(b)



Knowledge Discovery Illustration



A common relationship between leaf width and intercepted light per plant accounted for variations in width between fields, and for the difference between field and greenhouse

- Discover keys in numerical (RDF) data
 - **Keys:** combinations of properties that discriminate a resource
- Evaluate and understand their quality



- Experimental numerical data in wine flavour datasets (2011-2014)

How do we discriminate the wines??

PHIS

Plant and object information

lps-phis.supagro.inra.fr/phism3p/web/index.php?r=object%2Fview&type=Plant&id=http

Phenotyping Hybrid Information System *M3P*

Experimental Organization

Data

Tools

Phis Guest

Home / Plants / View / <http://www.phenome-fppn.fr/m3p/arch/2017/c17001684>

<http://www.phenome-fppn.fr/m3p/arch/2017/c17001684>



Showing 1-12 of 12 items.

Property	Value
http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.phenome-fppn.fr/vocabulary/m3p/2015#Plant
hasAlias	1692/CRAZI/ZM4112/WW/PSI_120/Loop4/ARCH2017-03-30
hasExperimentModalities	WW
hasGenotype	http://www.phenome-fppn.fr/m3p/g/CRAZI
hasRepetition	120
inExperiment	http://www.phenome-fppn.fr/m3p/ARCH2017-03-30
fromSpecies	http://www.phenome-fppn.fr/id/species/zeamays
hasInitialPosition	(Loop,4)
hasPotAlias	1692
isLocated	http://www.phenome-fppn.fr/m3p/phenoarch/looparea
fromSeedLotSample	http://www.phenome-fppn.fr/id/introduction/ZM4112
withinPot	http://www.phenome-fppn.fr/m3p/arch/2017/pc17000002184

Images

Showing 1-20 of 182 items.

#	Object URI	Parent URI	View Type	Date	Provider
1	m3p:arch/2017/c17001580535	m3p:arch/2017/c17001684 (1692/CRAZI/ZM4112/WW/PSI_120/Loop4/ARCH2017-03-30)	top0	2017-04-10 15:12:26.018512	elcom
2	m3p:arch/2017/c17001580536	m3p:arch/2017/c17001684 (1692/CRAZI/ZM4112/WW/PSI_120/Loop4/ARCH2017-03-30)	side0	2017-04-10 15:12:26.022512	elcom

PHIS

Environmental sensor

par07_p

lps-phis.supagro.inra.fr/phism3p/web/index.php?r=environmental-sensor%2Fview&id=

80%

Rechercher

Phenotyping Hybrid Information System M3P

Experimental Organization Data Tools This Guest

Sensor Alias	par07_p
URI	http://www.phenome-fppn.fr/m3p/arch/2013/sa130003
Sensor Type	RadiationSensor
Related Concept	http://purl.oclc.org/NET/ssnx/meteo/aws#QuantumSensor
Brand	Skye Instruments
Position (X,Y)	
Position (meter)	
Variable	PAR Light:weather station:micromole.m-2.s-1
In Service Date	2013-04-01
Date of Purchase	(not set)
Date of last Calibration	(not set)
Documents	Par_quantum_skp215.pdf

Data of par07_p

PAR Light:weather station:micromole.m-2.s-1

23 Nov 12:00 24 Nov 12:00 25 Nov 12:00 26 Nov 12:00 27 Nov 12:00 28 Nov 12:00 29 Nov 12:00 30 Nov

Periode of time

Week Month Range

2017-11-23 and 2017-11-30

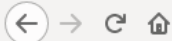
Refresh

© INRA MISTEA-LEPSE 2014-2017 (PHIS v.2.6 - 04th October 2017) ; © INRA MISTEA - SILEX

PHIS

Sensor information and references

W3 Ontology for Meteorolo X +



https://www.w3.org/2005/Incubator/ssn/ssnx/meteo/aws#QuantumSensor



Rechercher

QuantumSensor

measure the PAR directly in the range 0.4 to 0.7 micrometers

URI:	http://purl.oclc.org/NET/ssnx/meteo/aws#QuantumSensor
Label:	Quantum sensor
Source:	<i>skos:closeMatch</i> 'quantum' [GAMP 2.4.1.1]
Subclass of	aws:RadiationSensor and things that have a ssn:observes property who may be a dim:EnergyFlux
Paraphrase (experimental)	A aws:QuantumSensor is something that is a aws:RadiationSensor and has a ssn:observes property who may be a dim:EnergyFlux

Schema:

```
<owl:Class rdf:about="http://purl.oclc.org/NET/ssnx/meteo/aws#QuantumSensor">
  <rdfs:label>Quantum sensor</rdfs:label>
  <rdfs:comment>measure the PAR directly in the range 0.4 to 0.7 micrometers</rdfs:comment>
  <rdfs:subClassOf><owl:Class rdf:about="http://purl.oclc.org/NET/ssnx/meteo/aws#RadiationSensor"/>
</rdfs:subClassOf>
  <rdfs:subClassOf>
  <owl:Restriction>
    <owl:onProperty><owl:ObjectProperty rdf:about="http://purl.oclc.org/NET/ssnx/ssn#observes"/>
    </owl:onProperty>
    <owl:someValuesFrom><owl:Class rdf:about="http://purl.oclc.org/NET/ssnx/qu/dim#EnergyFlux"/>
    </owl:someValuesFrom>
  </owl:Restriction>
</rdfs:subClassOf>
  <dc:source>
    skos:closeMatch quantum sensor [GAMP 2.4.1.1]
    http://www.wmo.int/pages/prog/wcp/agm/gamp/gamp_en.html
  </dc:source>
</owl:Class>
```

PHIS

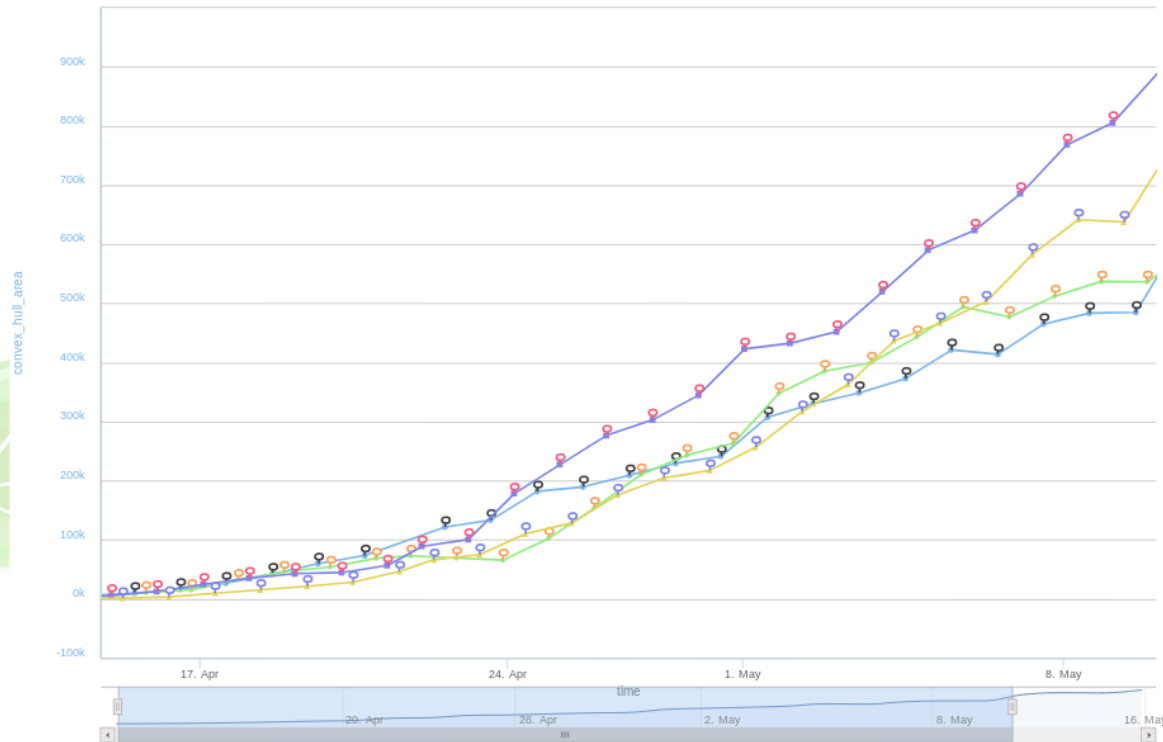
Provenance: Traits and associated images

147.99.24.182/phis-dev/web/index.php?r=graphic%2Fvisu&experiment=http%3A%2F%2F

Phenotyping Hybrid Information System M3P Experimental Organization Data Tools Pierre-Etienne Alary



Thursday, Apr 27, 07:49:48
 ● 0017/DZ_PG_20/ZM4344/WD/Veg_1/01_17/ARCH2017-03-30-Convex Hull Plant Area-side60-lepse: 210,512.50



● 0010/DZ_PG_19/ZM4367/WD/Veg_1/01_10/ARCH2017-03-30-Convex Hull Plant Area-side60-lep Images
 ● 0017/DZ_PG_20/ZM4344/WD/Veg_1/01_17/ARCH2017-03-30-Convex Hull Plant Area-side60-lep Images
 ● 0063/DZ_PG_41/ZM4378/WW/Rep_1/02_03/ARCH2017-03-30-Convex Hull Plant Area-side60-lep Images
 ● 0091/DZ_PG_18/ZM4373/WW/Rep_1/02_31/ARCH2017-03-30-Convex Hull Plant Area-side60-lep Images

PHIS

Event annotation

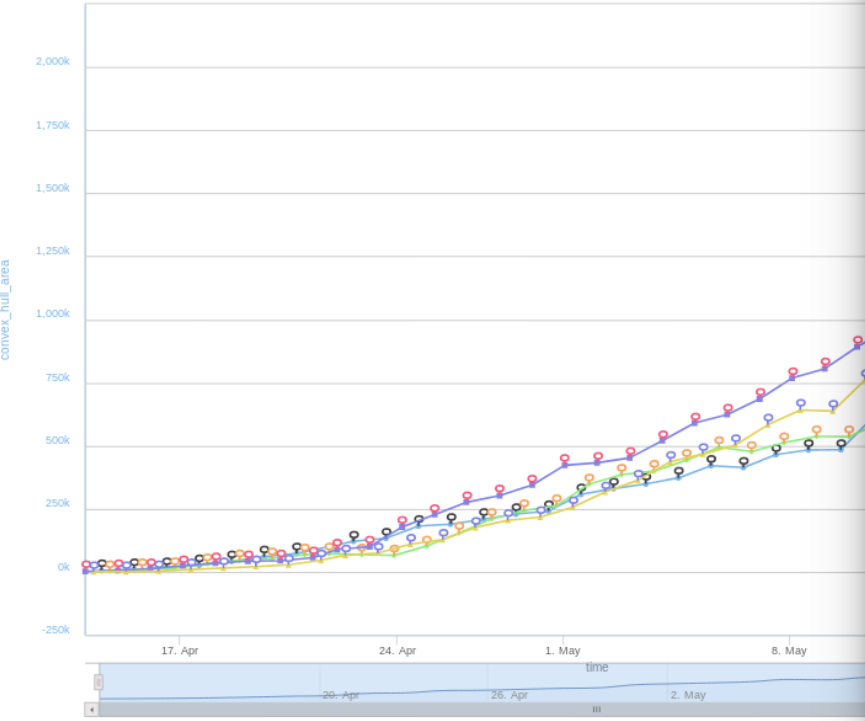
147.99.24.182/phis-dev/web/index.php?r=graphic%2Fvisu&experiment=http%3A%2F%2F 80 % Rechercher

Phenotyping Hybrid Information System M3P

Experimental Organization Data Tools Pierre-Etienne Alary



Friday, Apr 28, 04:27:01
 ● 0017/DZ_PG_20/ZM4344/WD/Veg_1/01_17/ARCH2017-03-30~Convex Hull Plant Area~side60~lepse: 243,378.00



- 0010/DZ_PG_19/ZM4367/WD/Veg_1/01_10/ARCH2017-03-30~Convex Hull Plant Area~side60~lepse: 243,378.00 Images
- 0017/DZ_PG_20/ZM4344/WD/Veg_1/01_17/ARCH2017-03-30~Convex Hull Plant Area~side60~lepse: 243,378.00 Images
- 0063/DZ_PG_41/ZM4378/WW/Rep_1/02_03/ARCH2017-03-30~Convex Hull Plant Area~side60~lepse: 243,378.00 Images
- 0091/DZ_PG_18/ZM4373/WW/Rep_1/02_31/ARCH2017-03-30~Convex Hull Plant Area~side60~lepse: 243,378.00 Images

PHIS - Mozilla Firefox

147.99.24.182/phis-dev/views/graphic/commu 80 %

Event or Expert Annotation

Author: pierre-etienne.alary@supagro.fr

IP: 10.146.2.250

Confidential: (oui)

Target (choose one): plant 0017/DZ_PG_20/ZM4344/WD/Veg_1/01_17/ARCH2017-03-30~Convex Hull Plant Area~side60~lepse: 243,378.00
 nearby image side60 2017-05-12T07:56:37+02:00

Datetime Event: 2017-05-12T07:45:07+02:00

Category:

Subject:

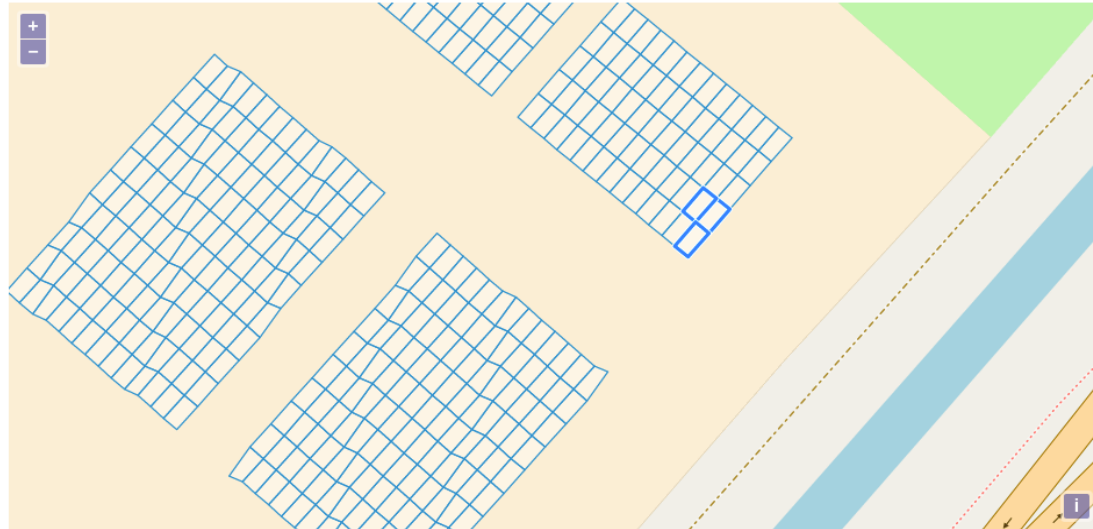
Content:

save

PHIS

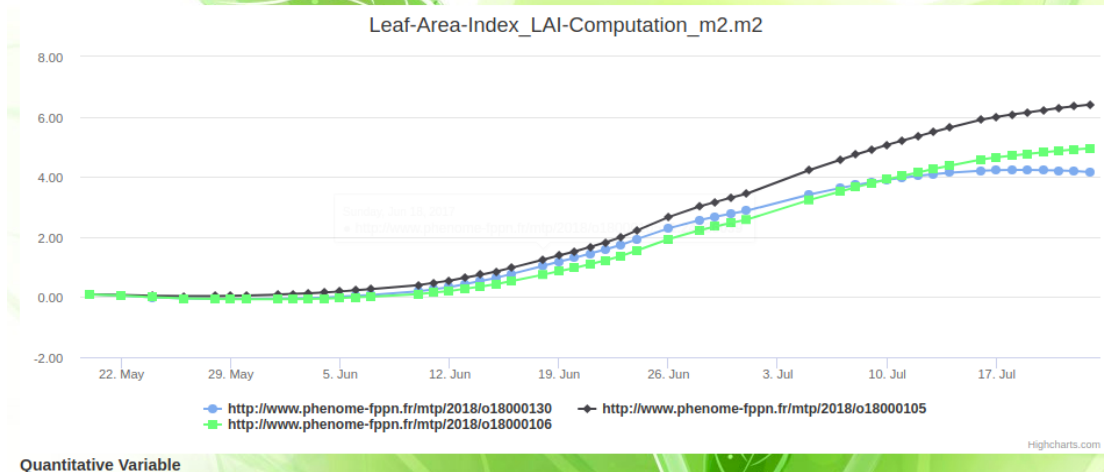
Data visualisation

<http://www.phenome-fppn.fr/diaphen/DIA2017-2>



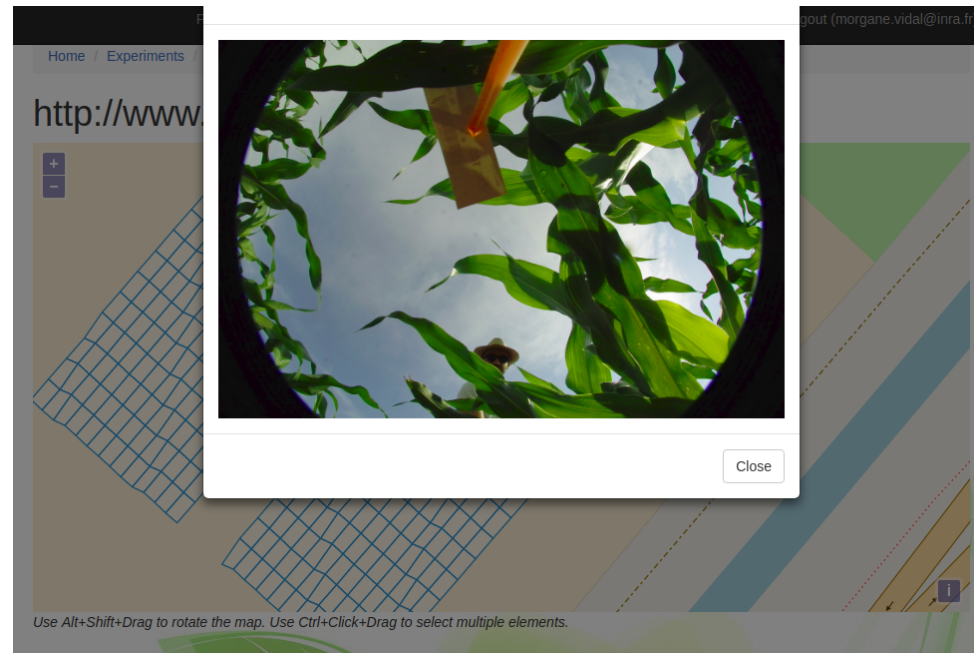
Use Alt+Shift+Drag to rotate the map. Use Ctrl+Click+Drag to select multiple elements.

Dataset(s) Visualization (On selected plot(s))





Trait – Images links





Create Variable

Variable Label *

MyNewTrait_MyNewMethod_NA

Trait

Trait label



Internal Label

MyNewTrait

Comment

This is a comment for y new trait, on which my new variable is focused.

Method

Method label



Internal Label

MyNewMethod

Comment

This is a comment for my new method, used to produce the values of my new variable.

Unit

Unit label



Ontologies References

In order to fill ontological references (URI) you can go to these ontologies :

- [AGROPORAL ?](#)
- [AGROVOC ?](#)
- [PLANT ONTOLOGY ?](#)
- [PLANTEOME ?](#)
- [CROP ONTOLOGY ?](#)
- [UNIT ONTOLOGY ?](#)

Related References

Entity	Relation	Reference URI	Hyperlink
Variable	skos:closeMatch		<input type="text"/> +
Variable	skos:narrower		<input type="text"/> x
Trait	skos:exactMatch		<input type="text"/> x
Method	skos:exactMatch		<input type="text"/> x

PHIS

Information searching

Search germplasms

Search Criteria

Species (ex.maize)

maize x

Germplasm

Select aliases:

Constraints (ex. leafArea < 0.4 & biovolume < 400) Add constraint

Constraint: 1 -

Variable *

leafArea (square meter)

Operator * **Value ***

< 0.4

Filter

<http://www.phenome-fppn.fr/m3p/g/iPG202>

[Return to the list](#)

Showing 1-4 of 4 items.

Property	Value
http://www.w3.org/1999/02/22-rdf-syntax-ns#type	http://www.phenome-fppn.fr/vocabulary/m3p/2015#Genotype
http://www.w3.org/2000/01/rdf-schema#label	iPG202
http://www.phenome-fppn.fr/vocabulary/m3p/2015#hasAlias	iPG202
http://www.phenome-fppn.fr/vocabulary/m3p/2015#fromSpecies	http://www.phenome-fppn.fr/id/species/zeamays

Object

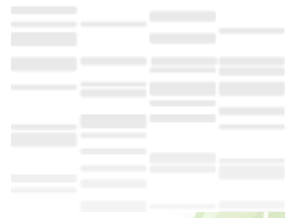
Plant validating search criteria

Showing 21-40 of 48 items.

URI	Object Alias
http://www.phenome-fppn.fr/m3p/2017/1/7001384	1384/DZ_PG_47/ZM4362/WD/Rep_6/24_04/ARCH2017-03-30
http://www.phenome-fppn.fr/m3p/2017/1/7001477	1477/DZ_PG_47/ZM4362/WW/Rep_TS_4/25_37/ARCH2017-03-30
http://www.phenome-fppn.fr/m3p/2017/1/7001506	1506/DZ_PG_47/ZM4362/WW/Rep_7/26_06/ARCH2017-03-30
http://www.phenome-fppn.fr/m3p/2017/1/7001676	1676/DZ_PG_47/ZM4362/WD/Rep_7/28_55/ARCH2017-03-30
http://www.phenome-fppn.fr/diaphen/2017/17000144	29/DZ_PG_47/ZM4362/WW/1/a/DIA2017-05-19
http://www.phenome-fppn.fr/diaphen/2017/17000145	29/DZ_PG_47/ZM4362/WW/1/b/DIA2017-05-19
http://www.phenome-fppn.fr/diaphen/2017/17000146	29/DZ_PG_47/ZM4362/WW/1/c/DIA2017-05-19
http://www.phenome-fppn.fr/diaphen/2017/17000147	29/DZ_PG_47/ZM4362/WW/1/d/DIA2017-05-19
http://www.phenome-fppn.fr/diaphen/2017/17000148	29/DZ_PG_47/ZM4362/WW/1/e/DIA2017-05-19

PHIS

Data analysis



Global Greenhouse Report

Name	Global Greenhouse Report
URI	http://www.phenome-fppn.fr/id/analysis/daglobal
Description	Visualization of a specified variable of an experiment. A HTML report is produced by this program.
Documents	

Experiment *

Trait *

View Label *

ARCH2017-03-30 - ZA17 experiment on objectSumArea parameter and side90 view

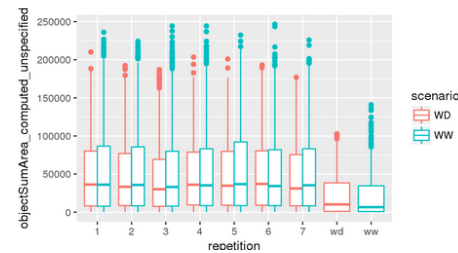
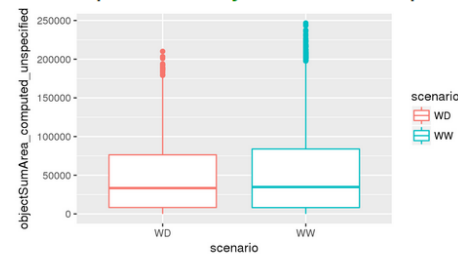
- 60 genotypes
- 2 scenarios: WD, WW
- 9 repetitions
- 1467 pots
- Scientist supervisors: Cabrera-Bosquet, Tardieu, Turc, Welcker
- Technical supervisors: Brichet, LUCHAIRE, Suard
- Experiment performed from 2017-03-30 to 2017-06-30
- Genotypes used in this experiment: IPG004, IPG007, IPG017, IPG026, IPG029, IPG062, IPG063, IPG066, IPG073, IPG077, IPG082, IPG089, IPG101, IPG103, IPG109, IPG110, IPG111, IPG116, IPG117, IPG119, IPG120, IPG121, IPG128, IPG131, IPG136, IPG138, IPG146, IPG148, IPG152, IPG153, IPG155, IPG158, IPG159, IPG164, IPG165, IPG167, IPG169, IPG173, IPG176, IPG181, IPG188, IPG189, IPG190, IPG194, IPG195, IPG202, IPG216, IPG228, IPG233, IPG234, IPG239, IPG303, IPG304, IPG310, IPG311, IPG312, IPG313, IPG314, IPG318, IPG321

Data Analysis

Showing 1-5 of 5 items.

#	Name ↓	URI	Description
1	Daily Greenhouse report	http://www.phenome-fppn.fr/id/analysis/dailyreportphis	Daily description of a PhenoArch experiment (imagery, environnement and so on...) running of it. A HTML report is produced by this program.
2	Environment Report	http://www.phenome-fppn.fr/id/analysis/daenvrfield	Description of the environment of a field experiment (meteo...). A HTML report program.
3	Environment Greenhouse Report	http://www.phenome-fppn.fr/id/analysis/daenvr	Description of environment of PhenoArch experiment (meteo...). A HTML report program.
4	Global Greenhouse Report	http://www.phenome-fppn.fr/id/analysis/daglobal	Visualization of a specified variable of an experiment. A HTML report is produced by th
5	Thermal Calculation Report	http://www.phenome-fppn.fr/id/analysis/dathermal	For a PhenoArch experiment, a thermal time is calculated according to the user's parent's metho). A HTML report and a csv file are produced by this program.

Description of objectSumArea parameter



PHIS

Workflow management

Home / Currents Tasks / Clean plant height using default



Clean plant height using default

Workflow name	Clean plant height using default
Start	09-01-2018 13:53
End	09-01-2018 14:29

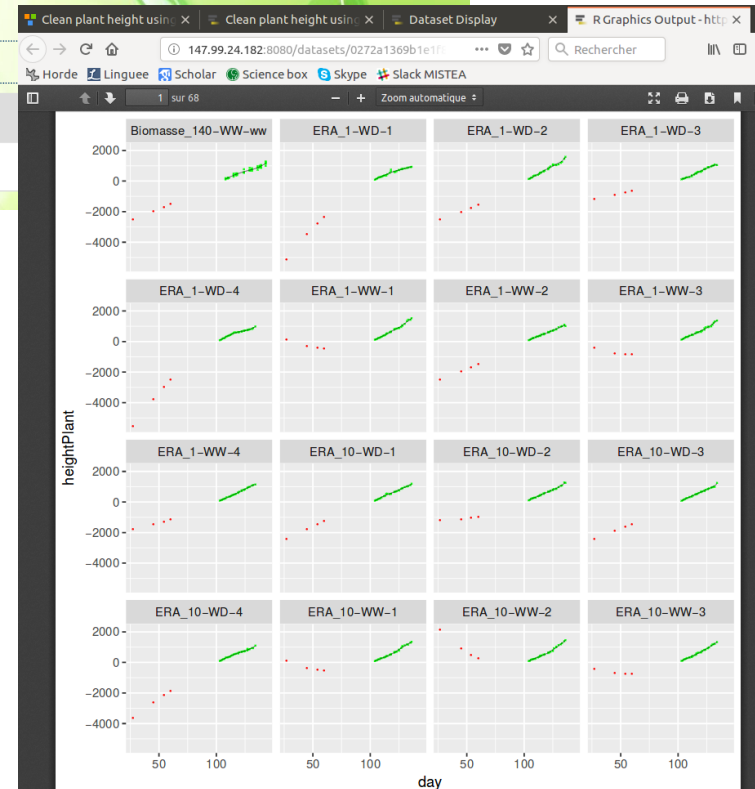
Open in Galaxy

Technical details

Invocation ID	40876639881ca029
History Id	6f91353f3eb0fa4a

The screenshot shows the Galaxy workflow interface. At the top, there are browser tabs for 'Clean plant height using default by Alray' and 'Dataset Display'. The main content area displays the workflow name and a list of steps:

- 11. Input: 2.8 MB, format: pdf, genome de référence: 2
- 10. Log from plotting: 9.3 KB
- 8. Lines data: ~100,000 lines, format: tabular, genome de référence: 2
- 7. Some data: 2.15 MB
- 6. log
- 5. treated data: 4.1 KB
- 4. log
- 3. plotam file: 2.15 MB
- 2. file
- 1. request file



PHIS

In short:

- ✓ Allows management of huge and complex data
 - Enables and facilitates cloud computing (data center, EGI)
 - distributed computing, distributed storage, backup
 - ✓ Open technologies
 - ✓ International identification (URI and DOI)
 - ✓ Semantic (ontologies, standardized vocabularies)
 - ✓ Portal interoperability and Open technologies
 - ✓ Provenance and reproducibility data processing
 - ✓ Flexible design
 - ✓ User right access
-
- ✓ 5 installations (field and greenhouse)
 - ✓ One instance → Over 300 Tb of data 25 plant species,

Conclusion and perspectives

- ✓ Maîtrise de l'organisation des données
- ✓ Outils de gestion et d'analyse partagés
- ✓ <http://www.phis.inra.fr/>
- ✓ Diffusion Emphasis (en cours) et plus largement
- ✓ <https://github.com/OpenSILEX>
- ✓ Faciliter et accompagner le déploiement + Formations
- ✓ Challenge méthodologique et culturel