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

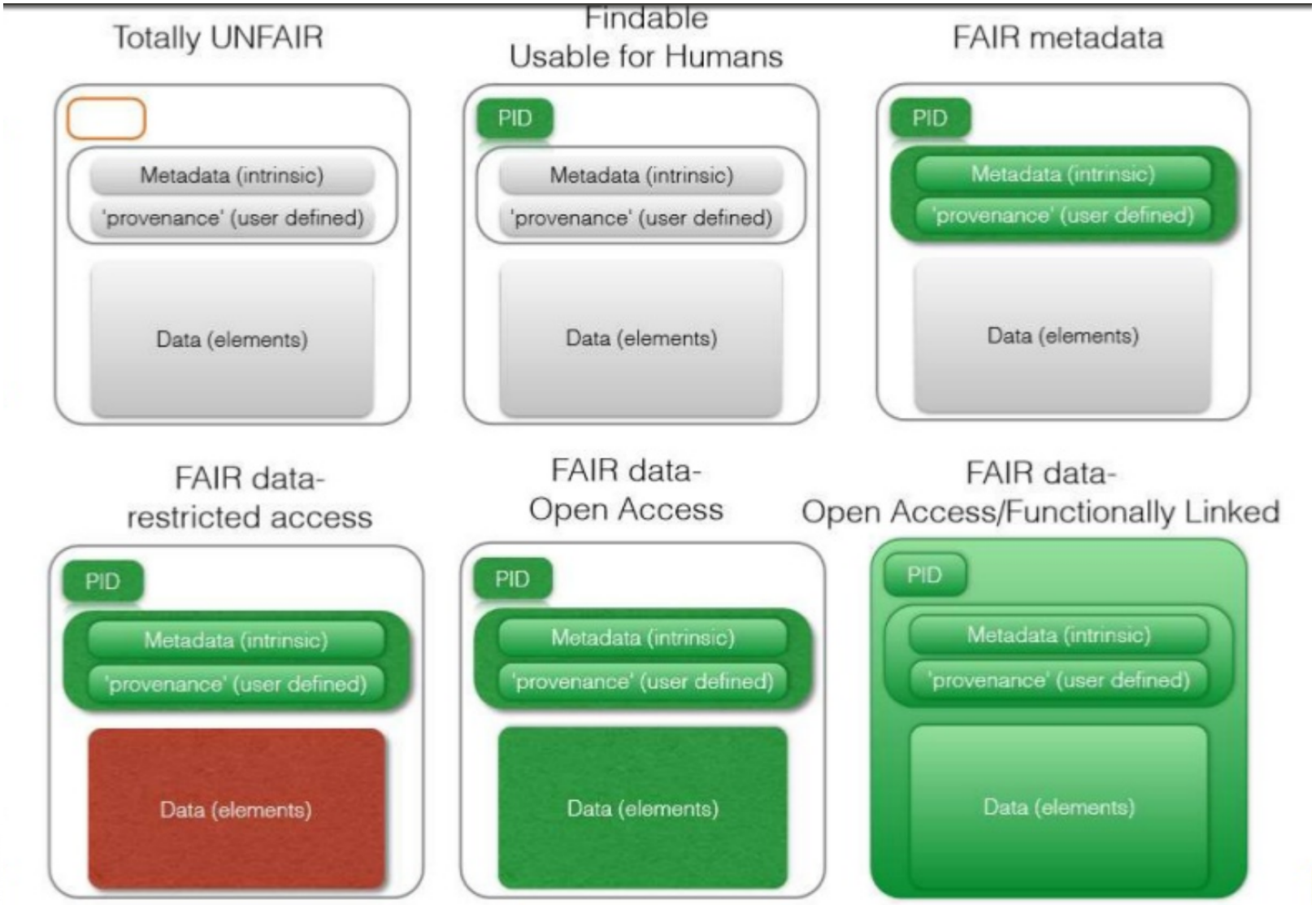
Accessible: open and standardized protocols (internet protocols), authentication* (if not open)

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

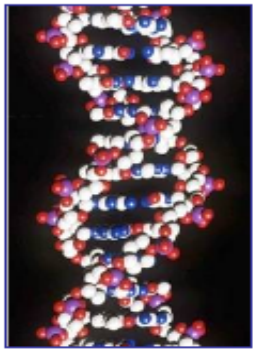
Reusable: provenance, domain relevant metadata for understanding

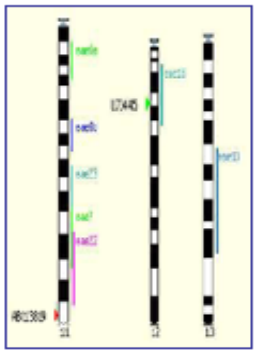
FAIR DATA

Findable, Accessible, Interoperable, Reusable



High Throughput Plant Phenotyping: searching for the most adapted genotypes





Many Plant Genotypes

Interactions

Various Environments

High frequency observations of trait dynamics for big set of Phenotypes

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

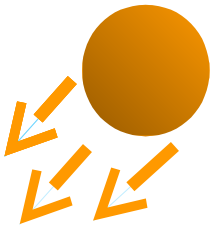
Emphasis e-infrastructure

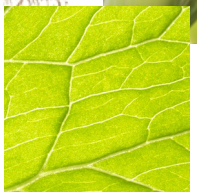
- Deals with several Petabytes of distributed data
- Makes FAIR data
- Based on Open technologies and standard (MIAPPE, BrAPI, etc)
- Standardized Identification
- Standardized Semantic
- Provenance and reproducibility data processing

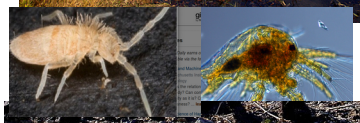
High Throughput Phenotypage data

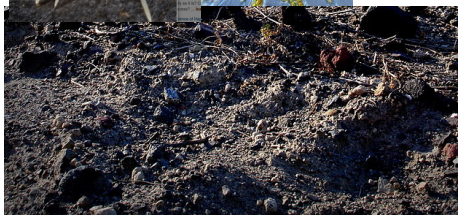
Different Scales

Interactions

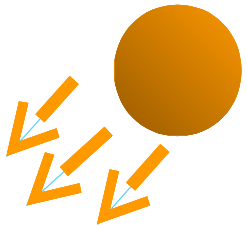








Different Species



Different Stages and Transformations

From various contexts

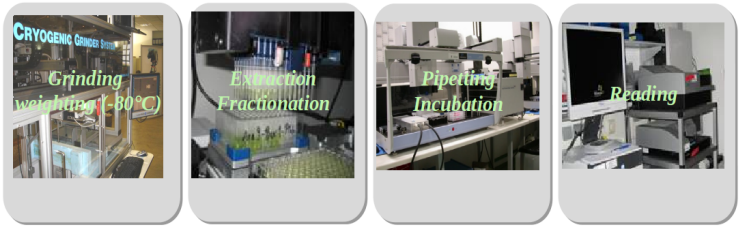
« omics » Platforms

Various data complex types

Genomics

Composition and the structure of biopolymers

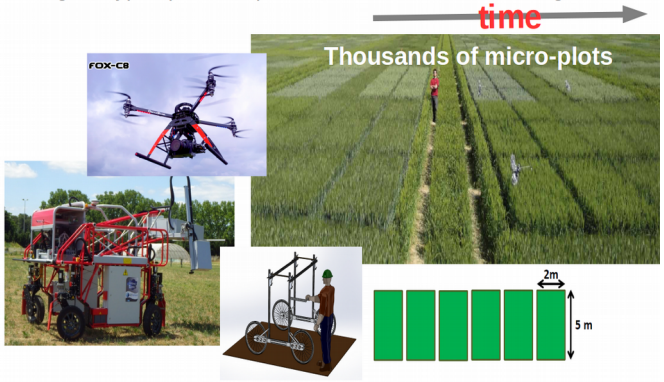
Quantification of metabolites and enzyme activities



Field Platforms

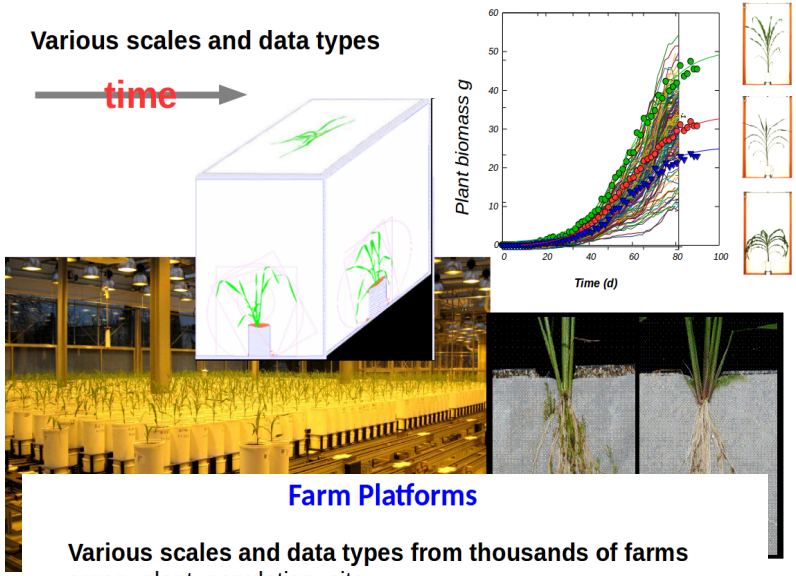
Various scales and data types

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



Green house Platforms

Various scales and data types

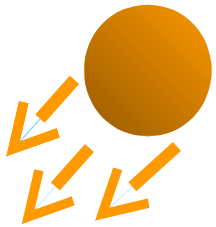


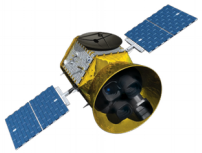
Farm Platforms

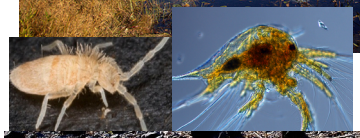
Various scales and data types from thousands of farms

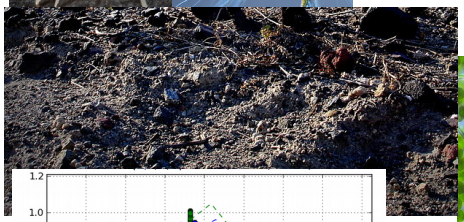
- organ, plant, population, site
- Images, sensors, human readings...

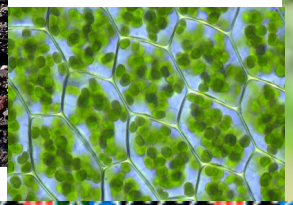
Heterogenous Data Sources

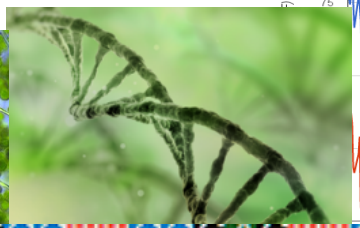


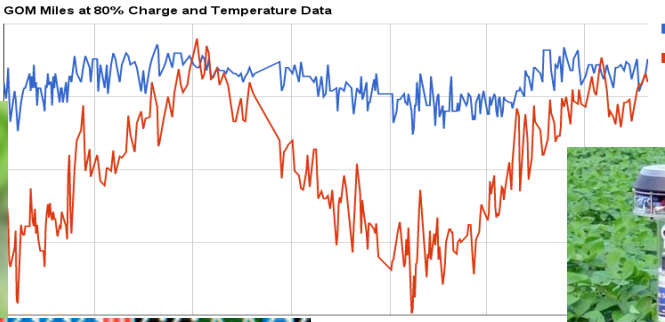


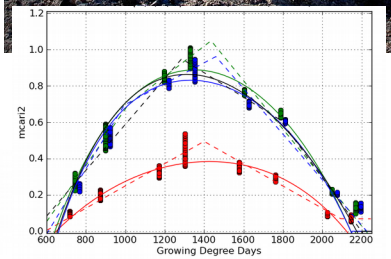




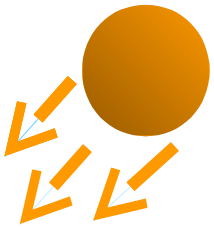


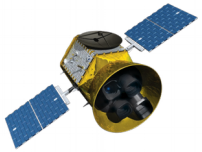




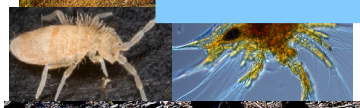


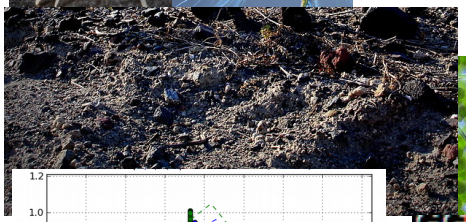
Heterogenous Data Sources

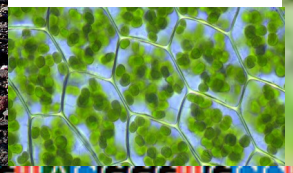


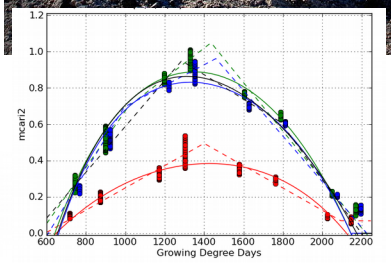


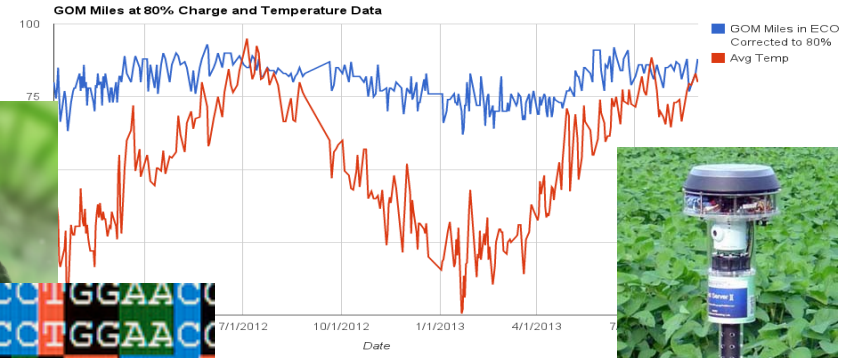
Intensive data integration needs...



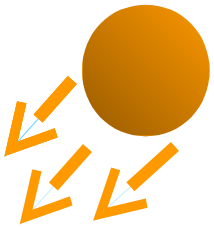


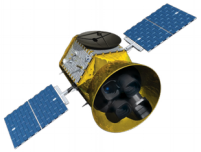




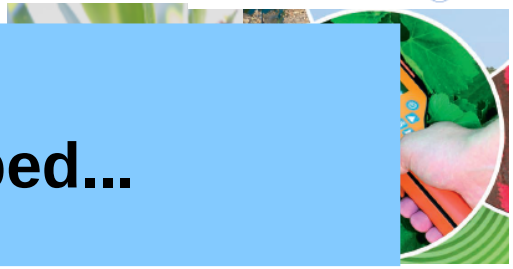


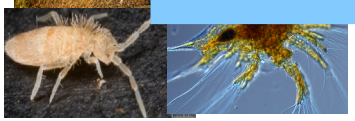
Heterogenous Data Sources

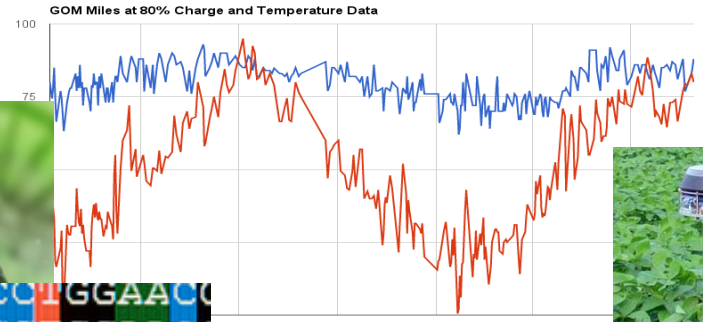


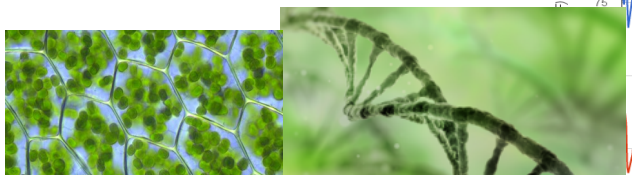


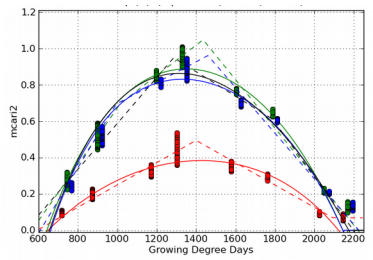
- **Orphan data → Worthless!**
- **Data has value if they are grouped...**











How to structure data?

PHIS: An Information System for High Throughput Phenotyping

- Designed and developped 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?

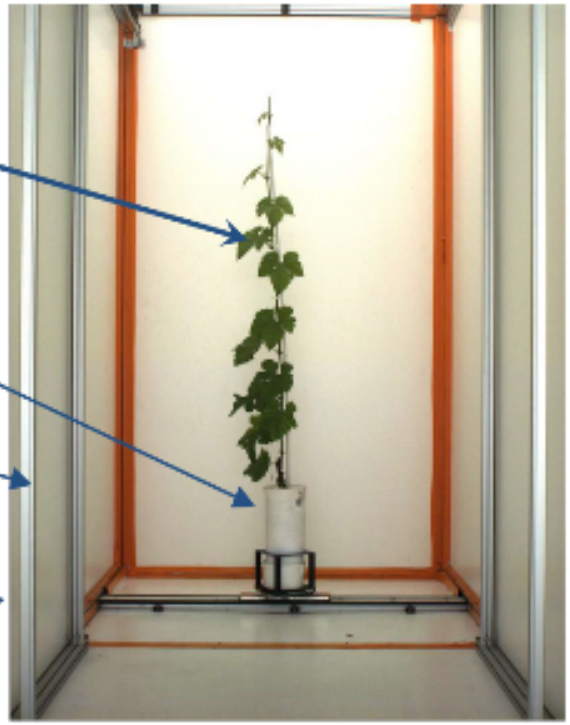
URI of plant :
mp3:arch/2014/pl/000000012

URI of pot :
mp3:arch/2001/pt/000001542

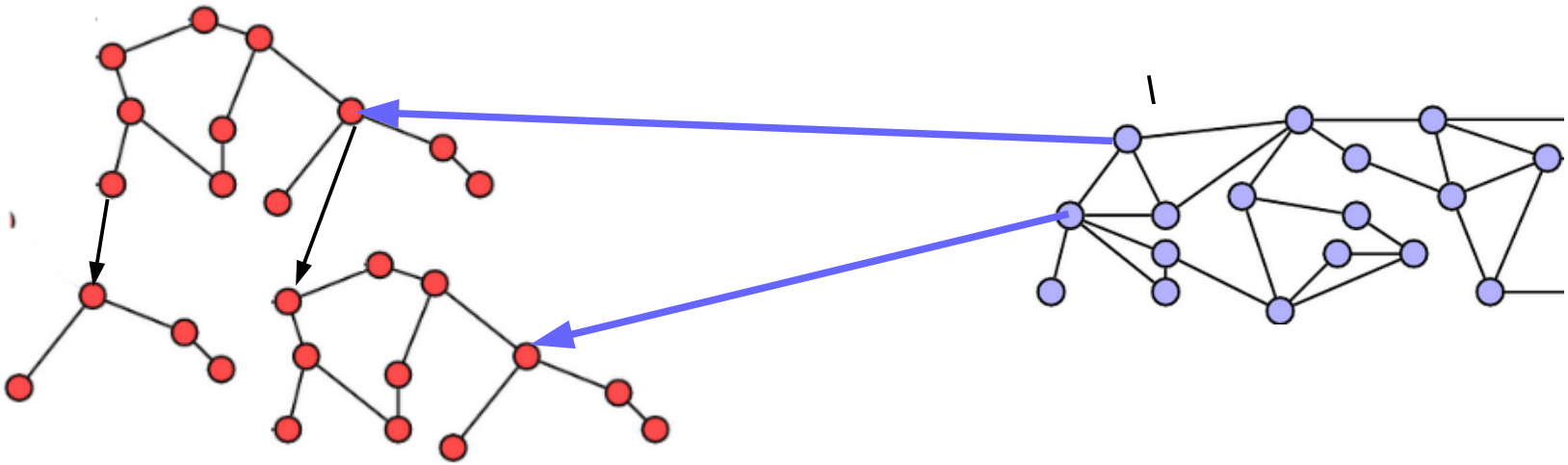
URI of cabin :
mp3:arch/2010/ca/cabine2

URI of camera :
mp3:arch/2011/ss/00003312

URI of image :
mp3:arch/2015/im/000000564



Semantic Ressources



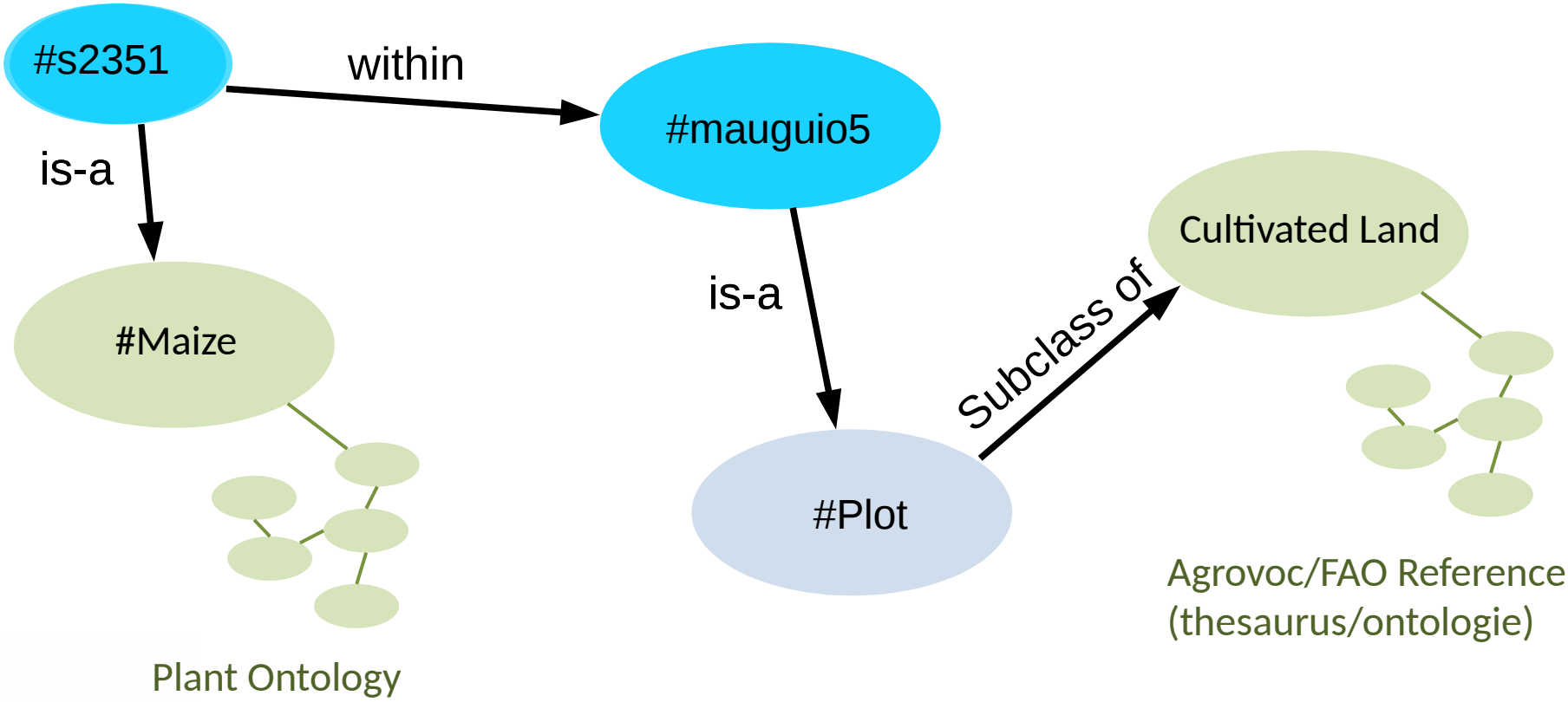
Reference ontologies

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

Application ontologies
Web Annotation Ontology

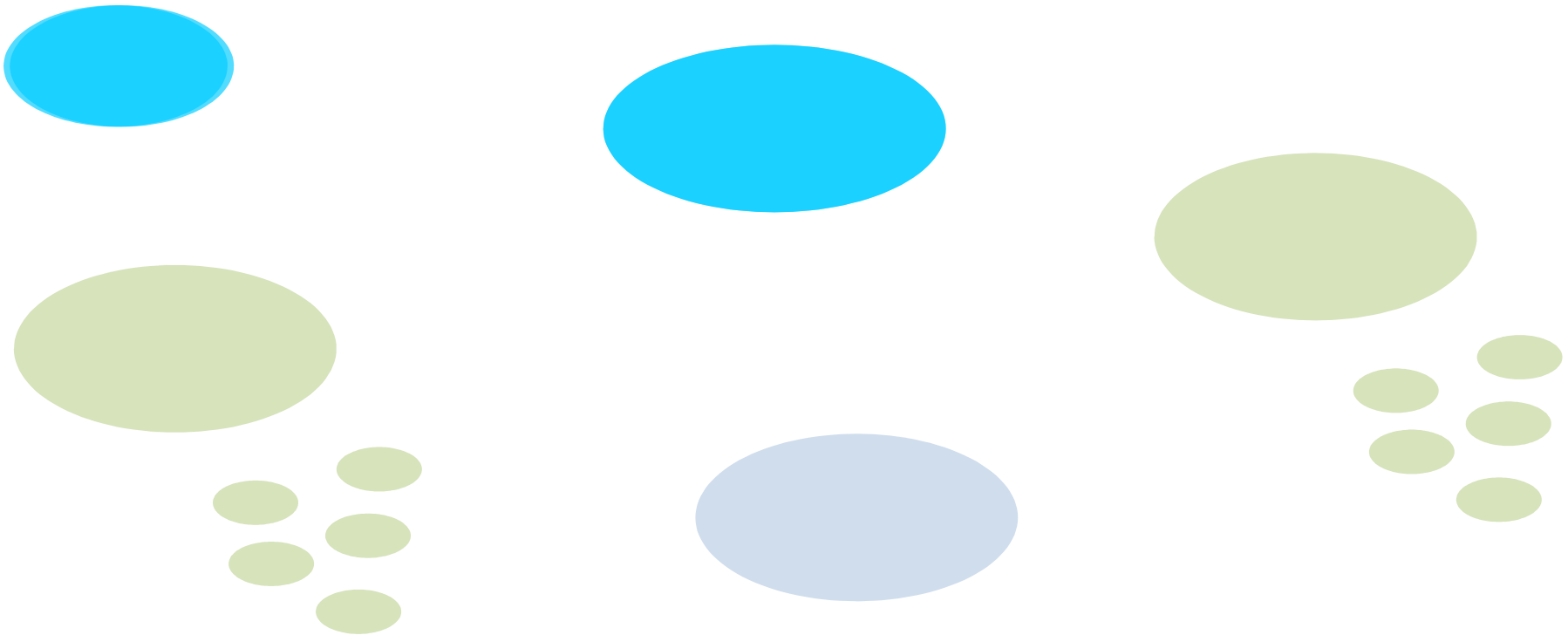
Semantics for the Structuring of Data

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

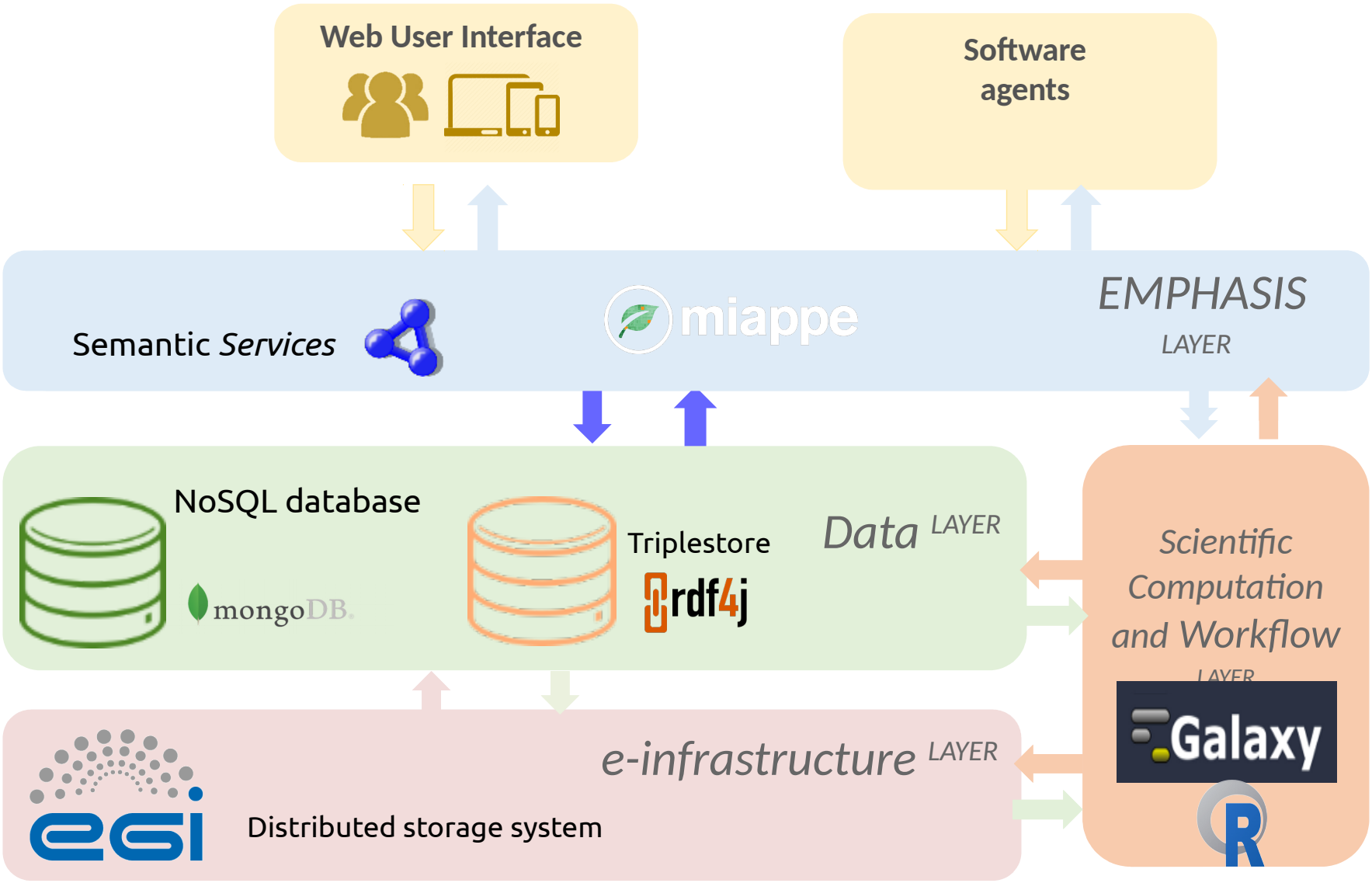


Semantics for the Structuring of Data

Metadata / Ontologies provide the meaning of the data



PHIS Architecture



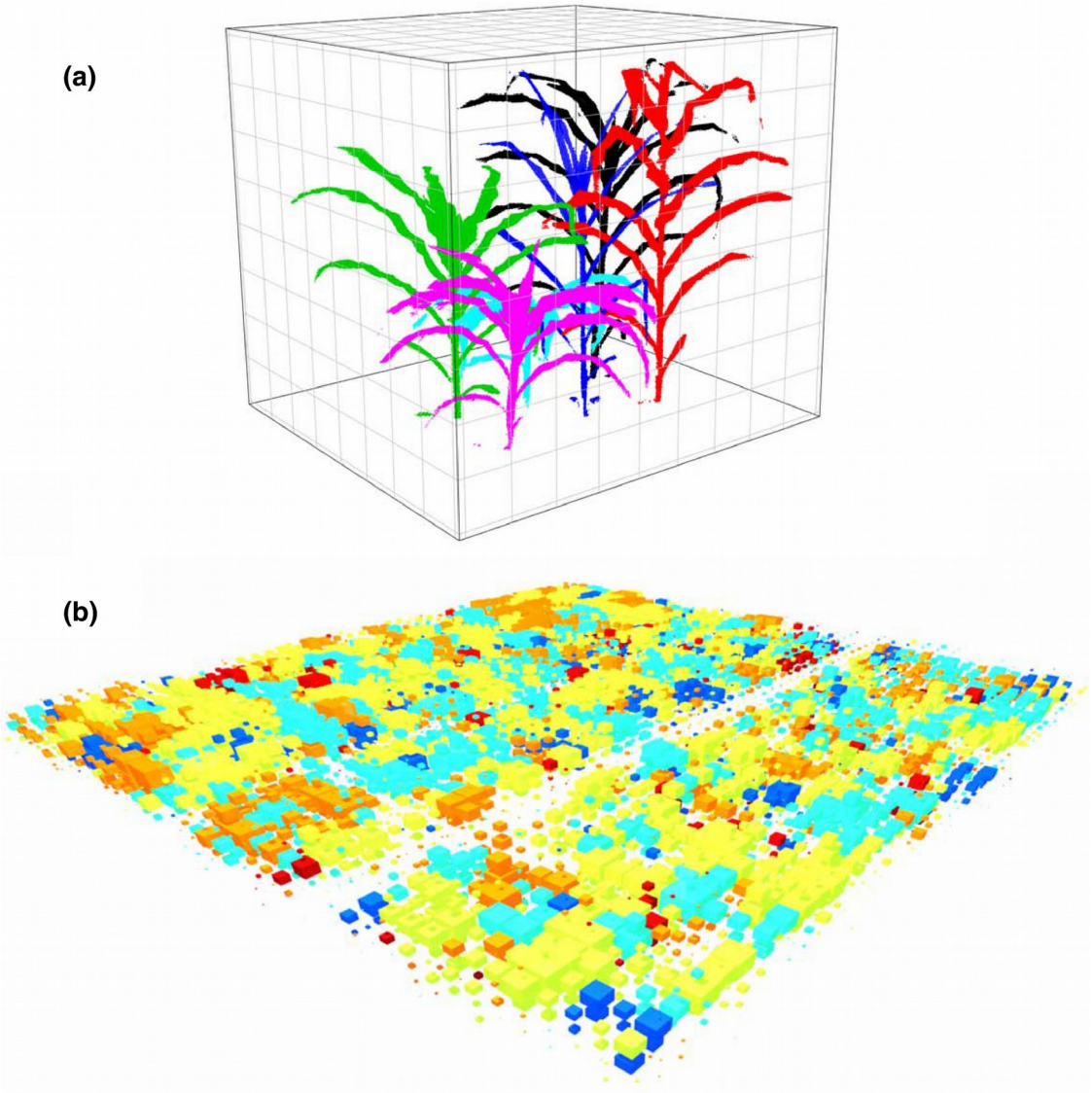
PHIS

Make easier Cloud computing

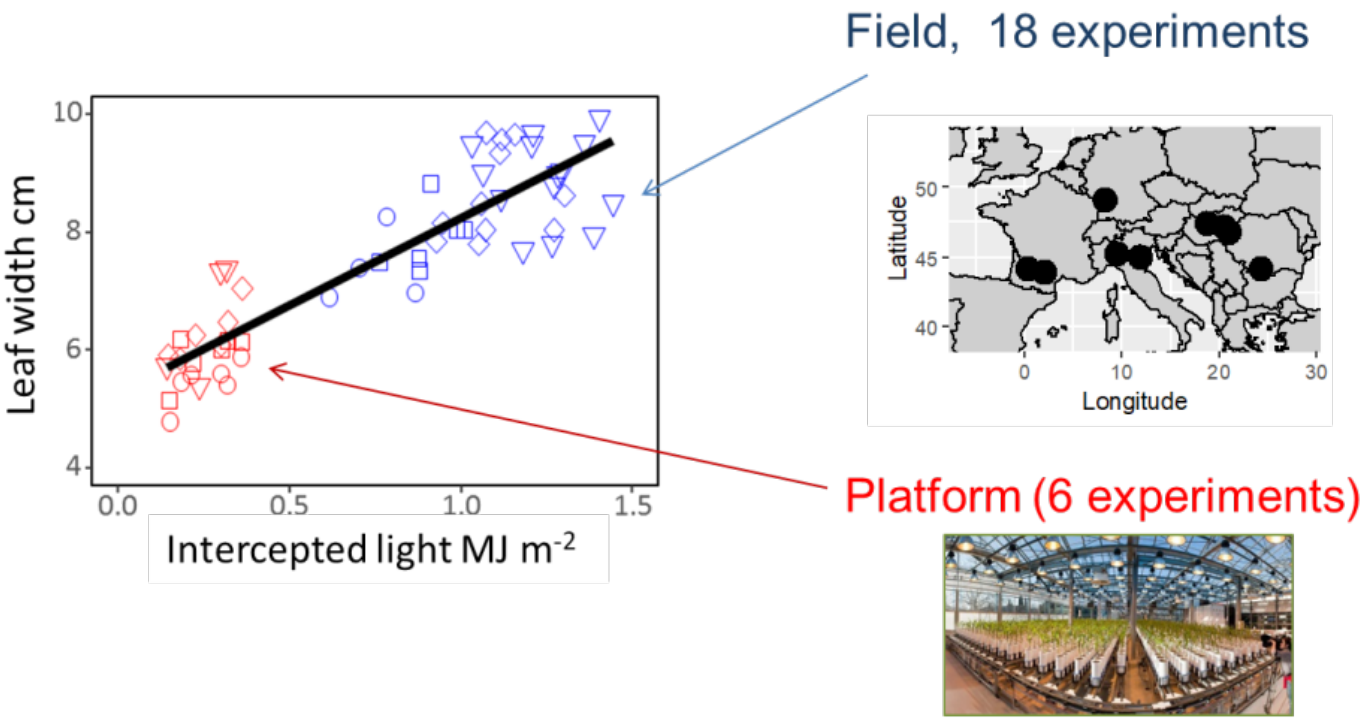
- **IRODS / ONeData**
  (distributed storage of image datasets)
- **France Grille, EGI**
- **INRA Data Center**

Knowledge Discovery Illustration

PHIS provides contextualisation: intercepted light value



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

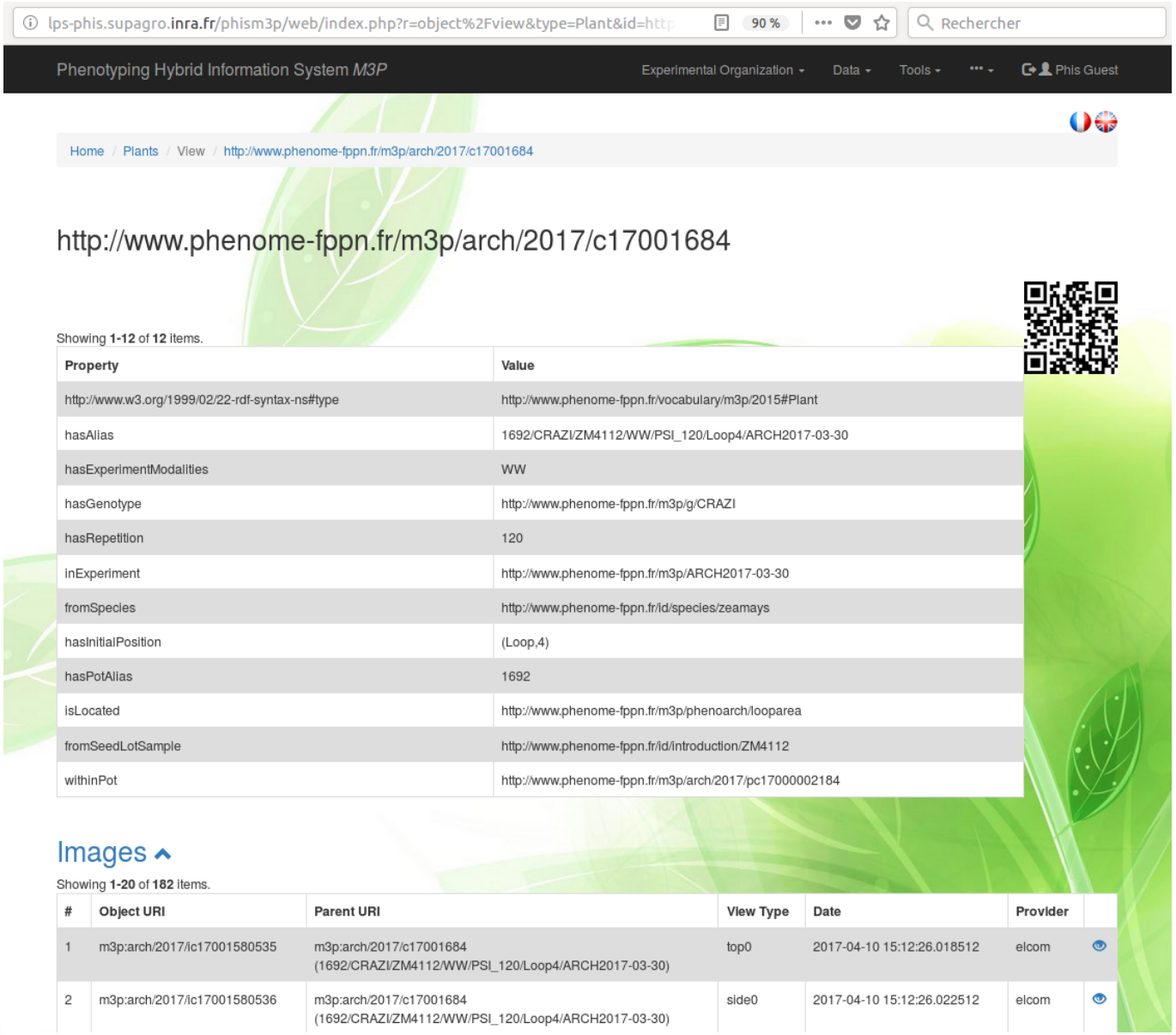
- Discover keys in numerical (RDF) data
  - **Keys:** combinations of properties that discriminate a resource

- Evaluate and understand their quality

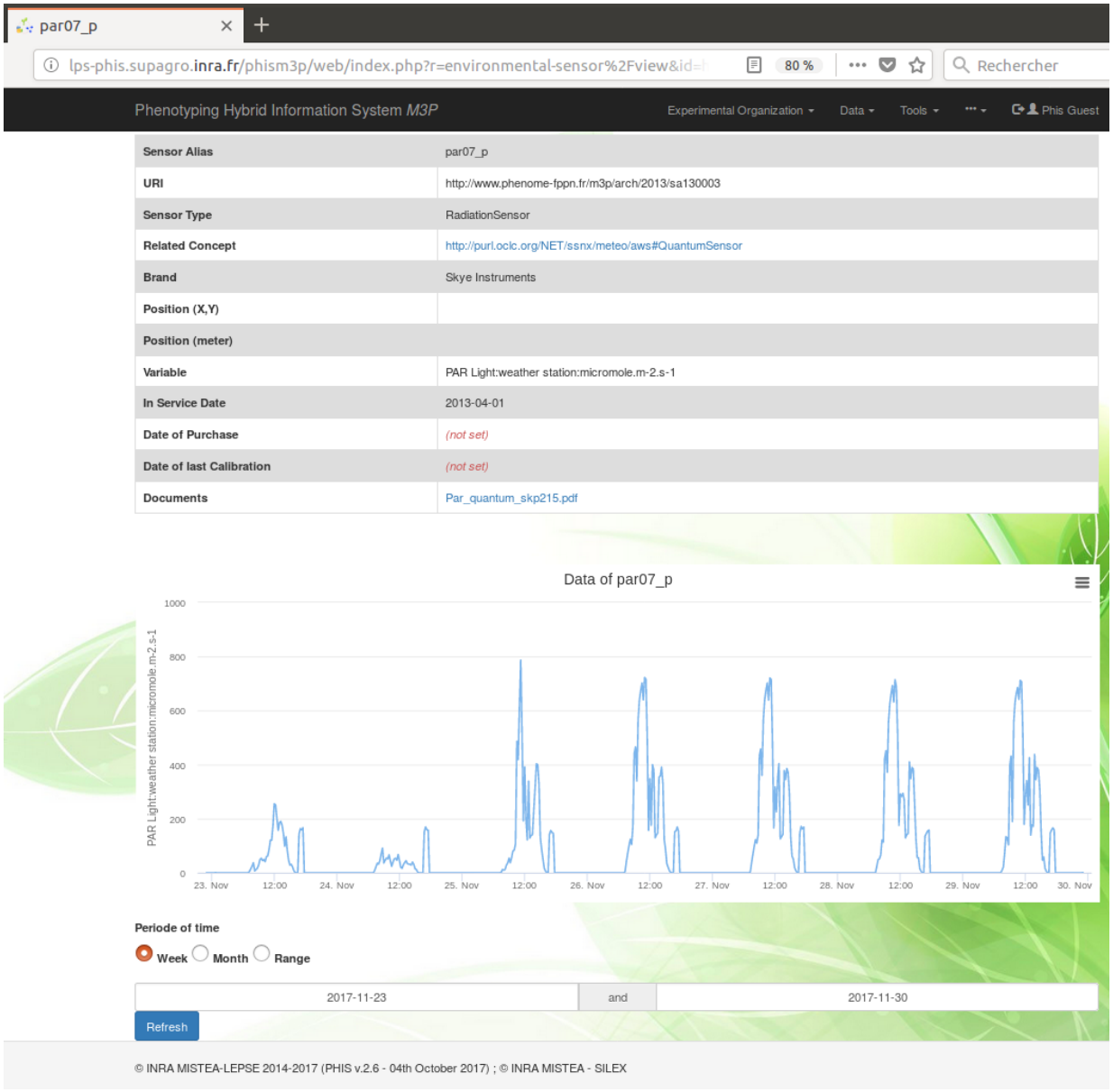
- Experimental numerical data in wine flavour datasets (2011-2014)

How do we discriminate the wines??

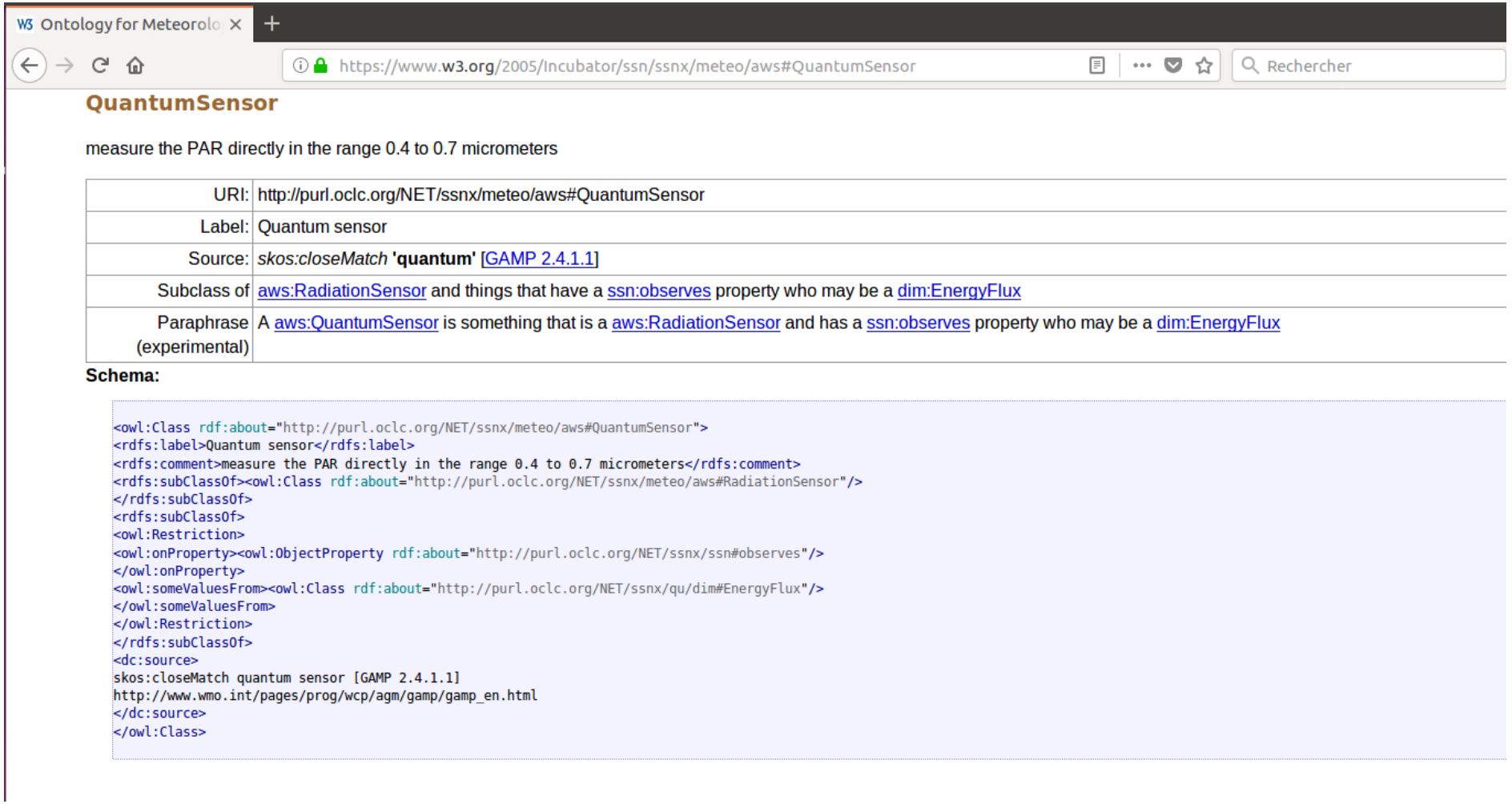
PHIS
Plant and object information



PHIS
Environmental sensor

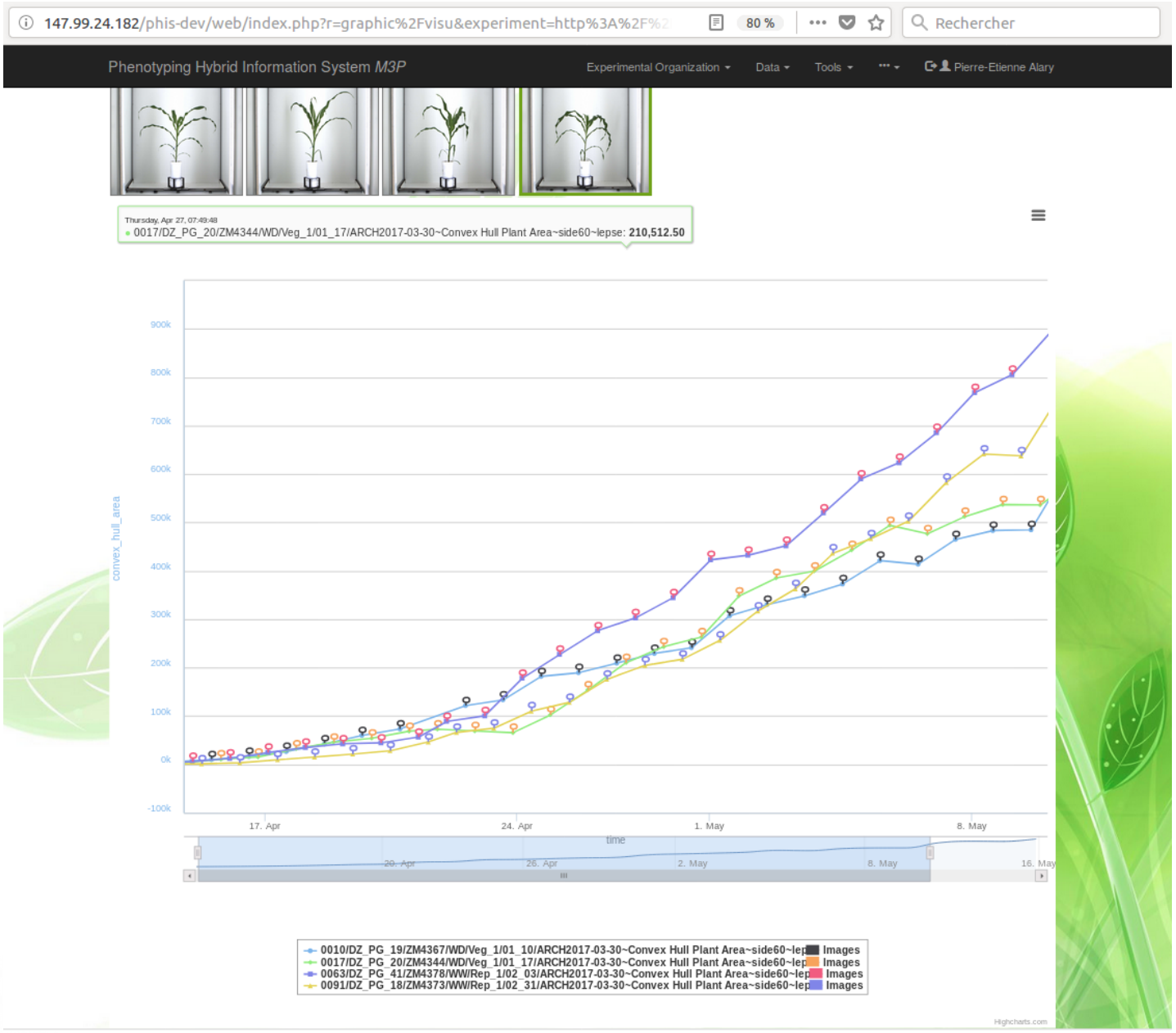


PHIS
Sensor information and references

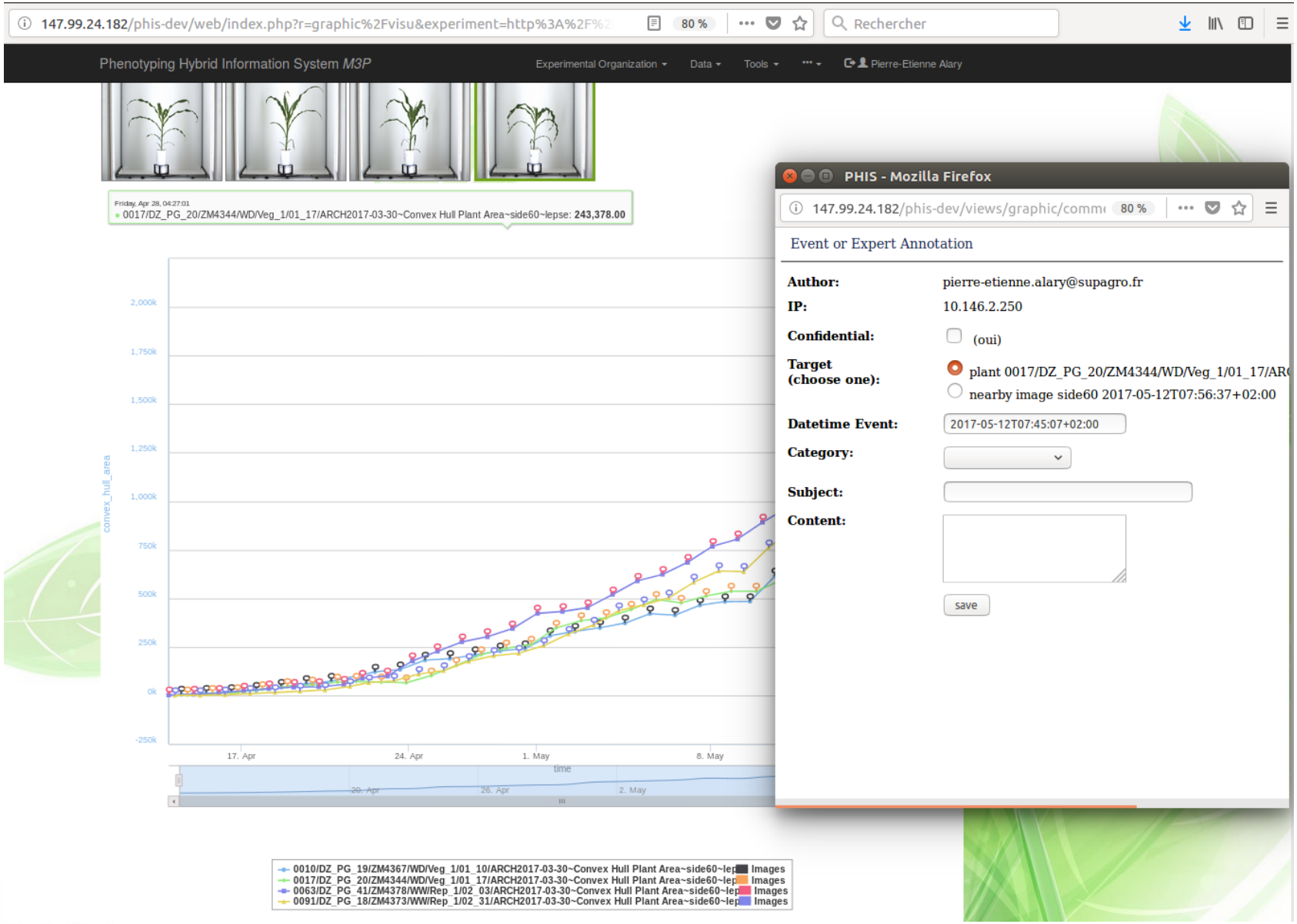


PHIS

Provenance: Traits and associated images

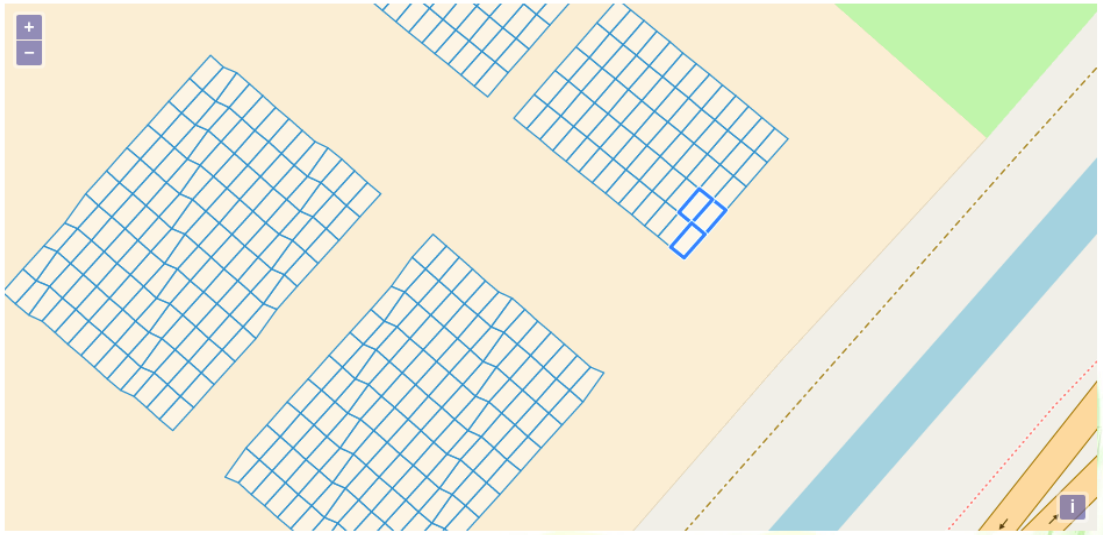


PHIS
Event annotation



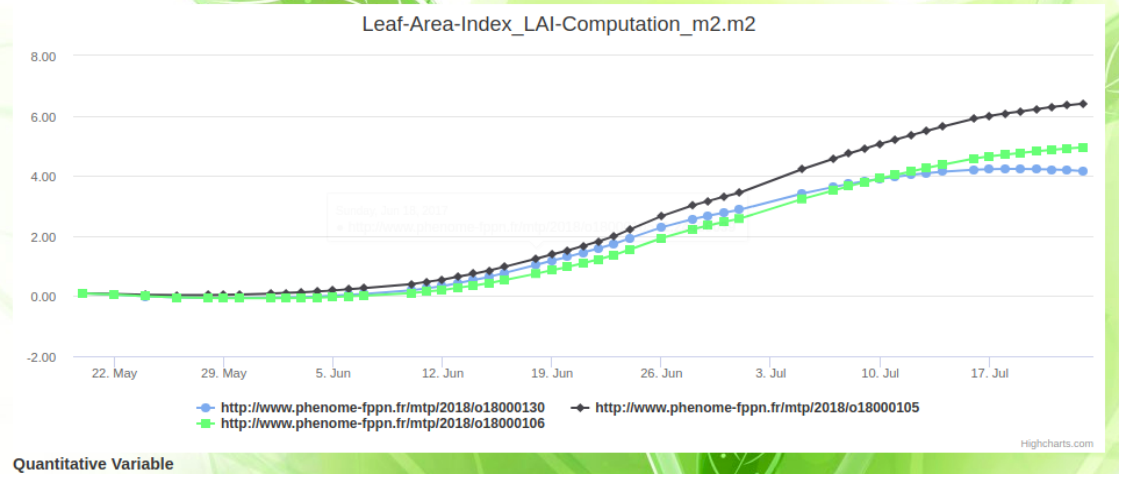
PHIS
Data visualisation

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



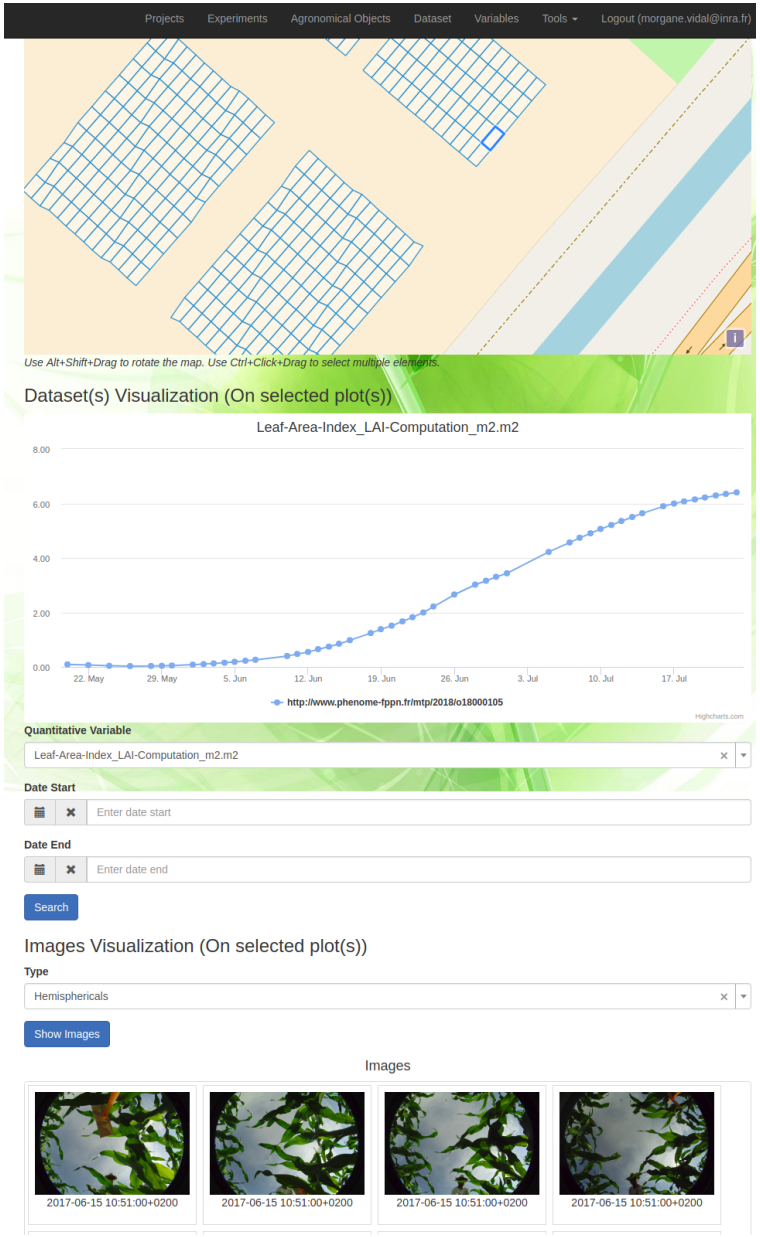
Use Alt+Shift+Drag to rotate the map. Use Ctrl+Click+Drag to select multiple elements.

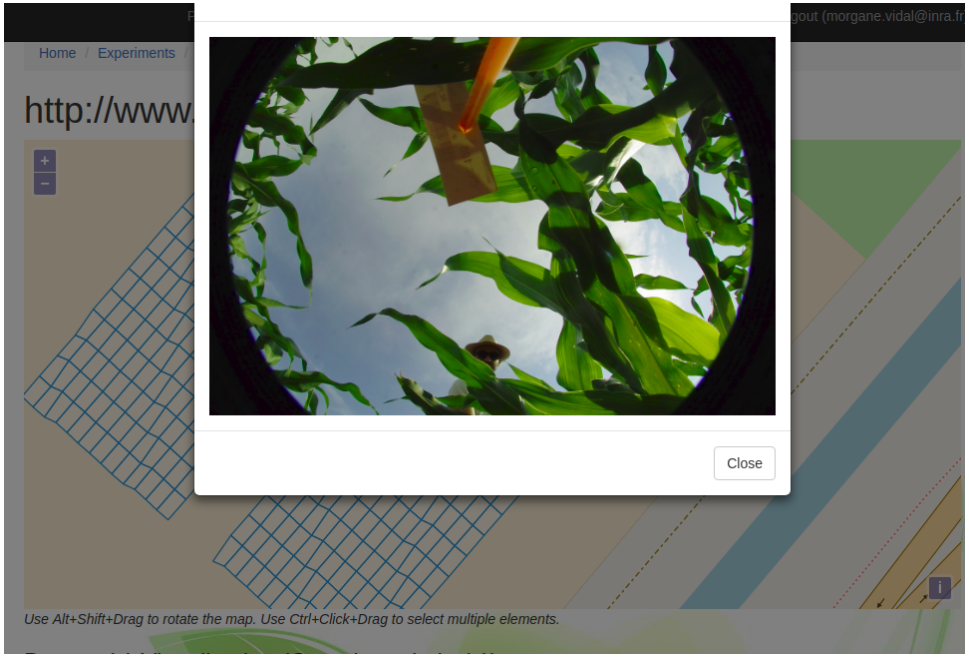
Dataset(s) Visualization (On selected plot(s))



Quantitative Variable

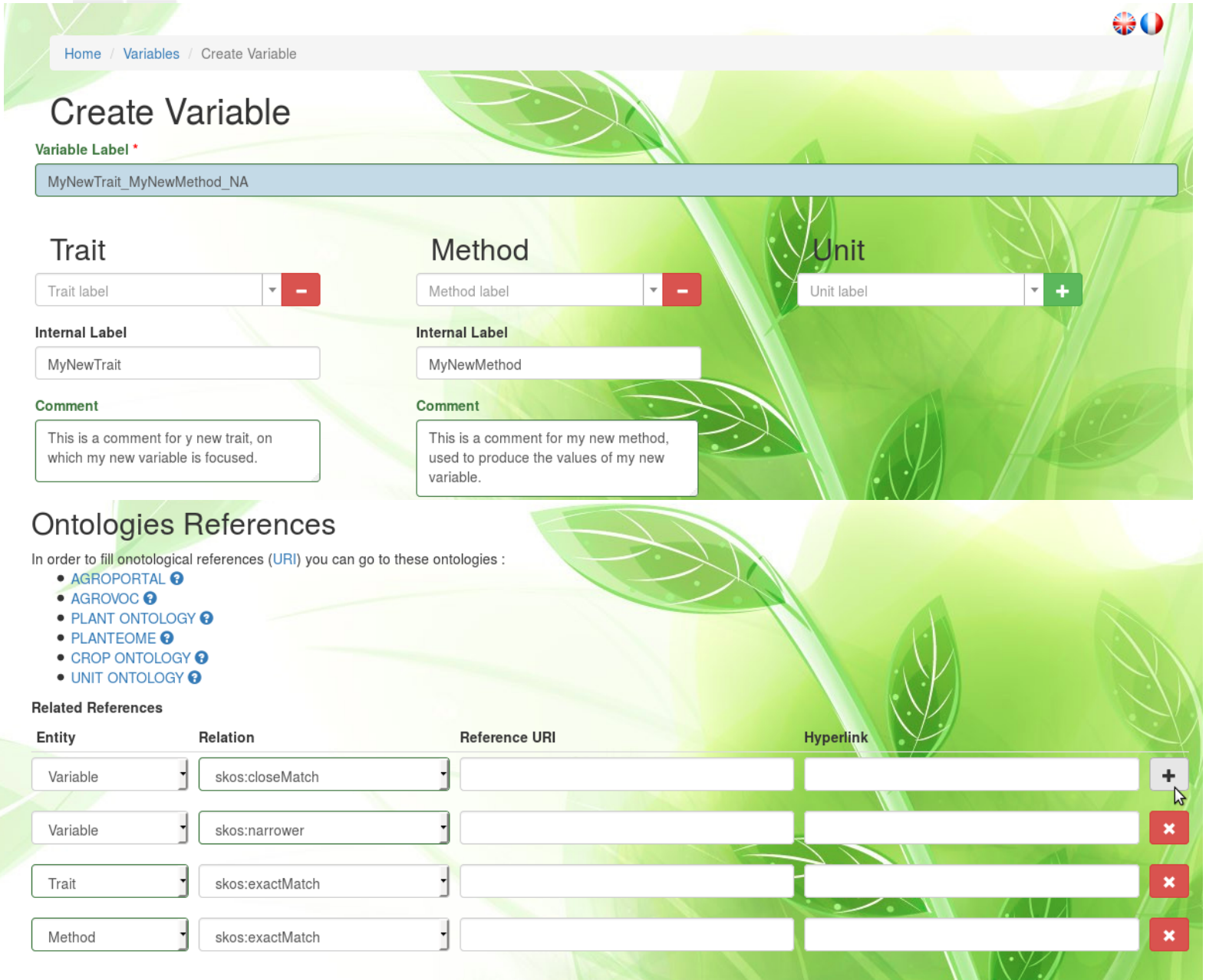
Trait – Images links



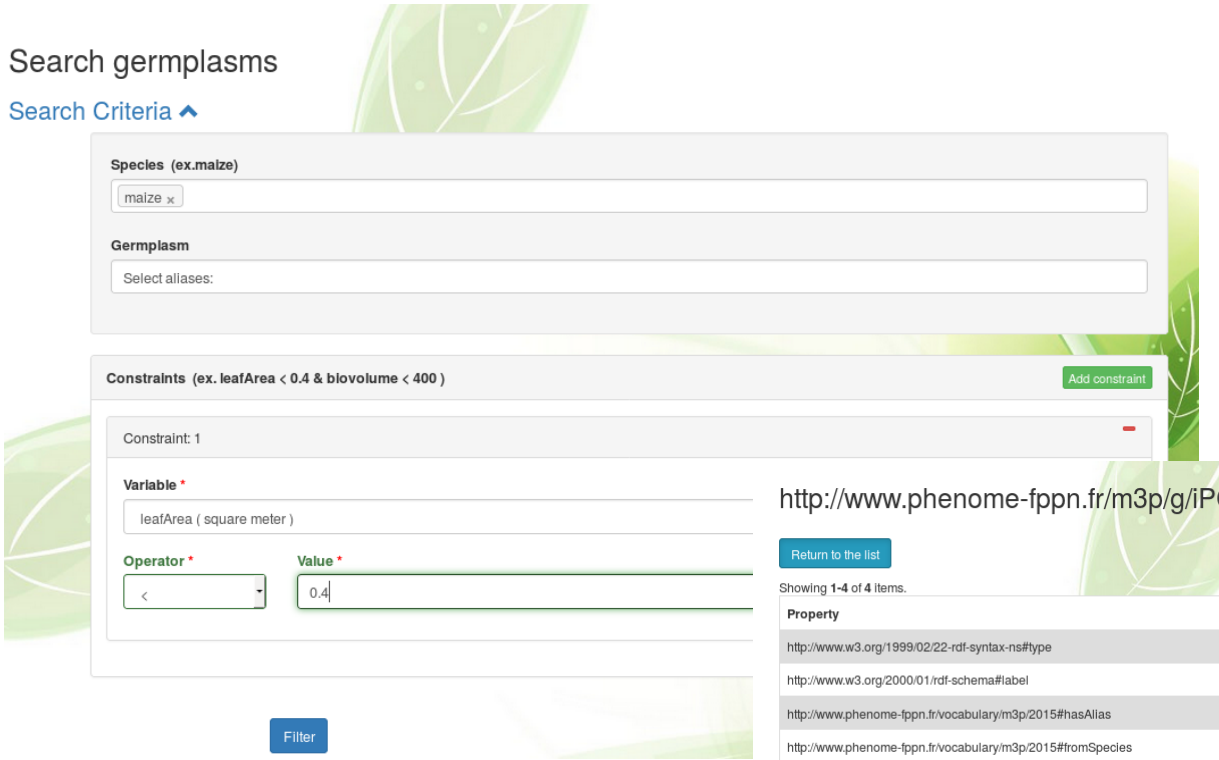


PHIS

Variables management and intetoperability



PHIS
Information searching



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/arch/2017/c17001384	1384/DZ_PG_47/ZM4362/WD/Rep_6/24_04/ARCH2017-03-30
http://www.phenome-fppn.fr/m3p/arch/2017/c17001477	1477/DZ_PG_47/ZM4362/WW/Rep_TS_4/25_37/ARCH2017-03-30
http://www.phenome-fppn.fr/m3p/arch/2017/c17001506	1506/DZ_PG_47/ZM4362/WW/Rep_7/26_06/ARCH2017-03-30
http://www.phenome-fppn.fr/m3p/arch/2017/c17001675	1675/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

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 *
ARCH2017-03-30

Trait *
objectSumArea

View Label *
side90

Run

Data Analysis

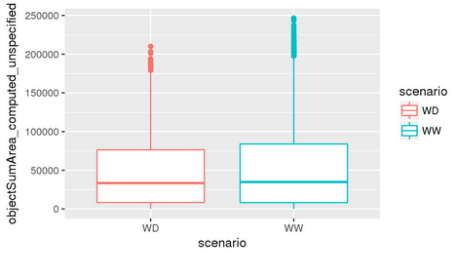
Showing **1-5** of **5** items.

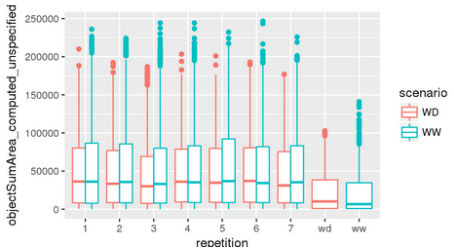
#	Name	URI	Description
1	Daily Greenhouse report	http://www.phenome-fppn.fr/id/analysis/dailyreportphis	Daily description of a PhenoArch experiment (imagery, environnement and so on...) running of it. A HTML report is produced by this program.
2	Environment Field Report	http://www.phenome-fppn.fr/id/analysis/daenvirfield	Description of the environment of a field experiment (meteo...). A HTML report i program.
3	Environment Greenhouse Report	http://www.phenome-fppn.fr/id/analysis/daenvir	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 t
5	Thermal Time Calculation Report	http://www.phenome-fppn.fr/id/analysis/dathermal	For a PhenoArch experiment, a thermal time is calulated according to the user's parent's metho). A HTML report and a csv file are produced by this program.

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

Description of objectSumArea parameter





PHIS
Workflow management

Home / Currents Tasks / Clean plant height using default

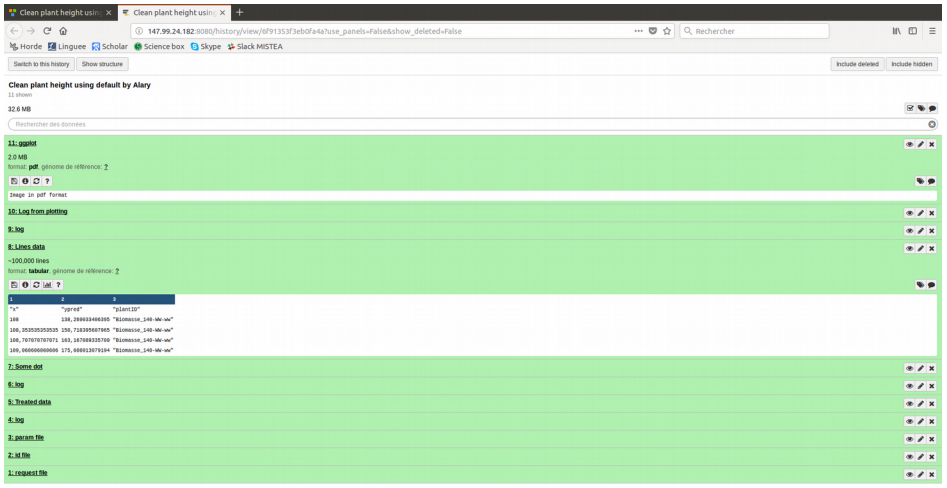
Clean plant height using default

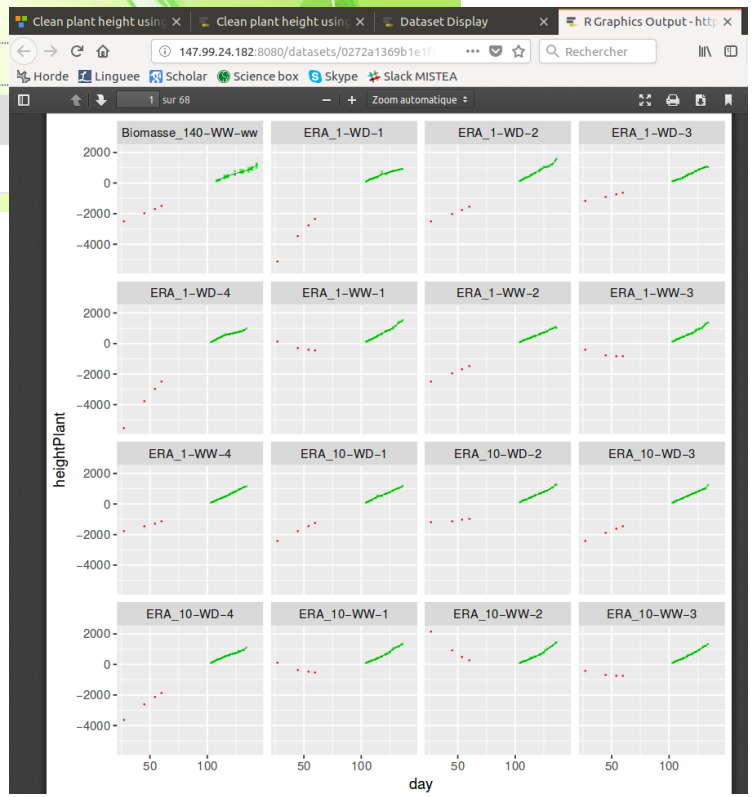
Workflow name	Clean plant height using default
Start	09-01-2018 13:53
End	09-01-2018 14:29

Open in Galaxy

Technical details

Invocation ID	40876639881ca029
History Id	6f91353f3eb0fa4a





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