# Systèmes d'information et analyses de données

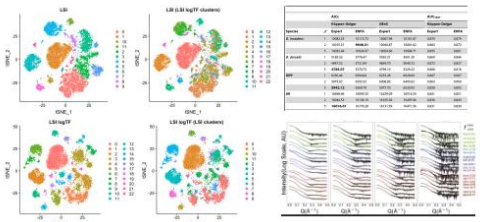
**mardi 05 avril, Paris**

Isabelle Alic, Pascal Neveu, Cyril Pommier

ANR11-INBS-0012

# Phenomics Data Challenge

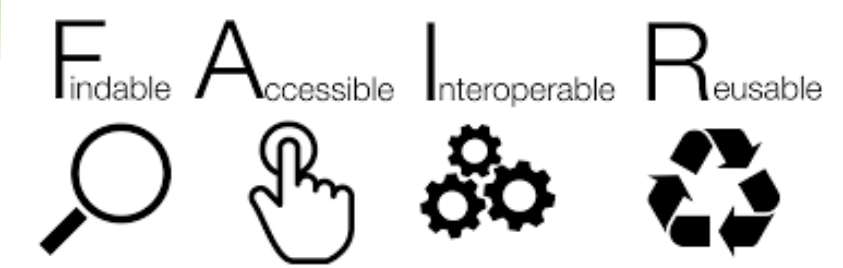
- **Experimentations in the domain of phenomics**
  - Expensive, require a lot of resources and often very hard
  - Cannot be reproduced
  - Huge and **very complex datasets**



- Strong needs of **transparence** and reproducibility

- **Give value to data:** re-analyses, meta-analyses and new analyses
  → impossible without **advanced data management**

# FAIR Principles




**Findable:** **persistent ID**, indexed in portals, standardized and relevant metadata

***Challenge: coordinated and sustainable data services***

**Accessible:** open and standardized protocols, **license rights**

***Challenge: cultural evolution of teams***

**Interoperable (technology, syntax, semantic):** shared standardized formats, vocabularies and **formal languages for knowledge representation,**

***Challenge: skill developement (interdisciplinary)***

**Reusable:** **provenance**, relevant metadata for understanding **across disciplines,**

***Challenge: new analysis methods***

# Complex Data Challenge

- A need for new generation of Information Systems

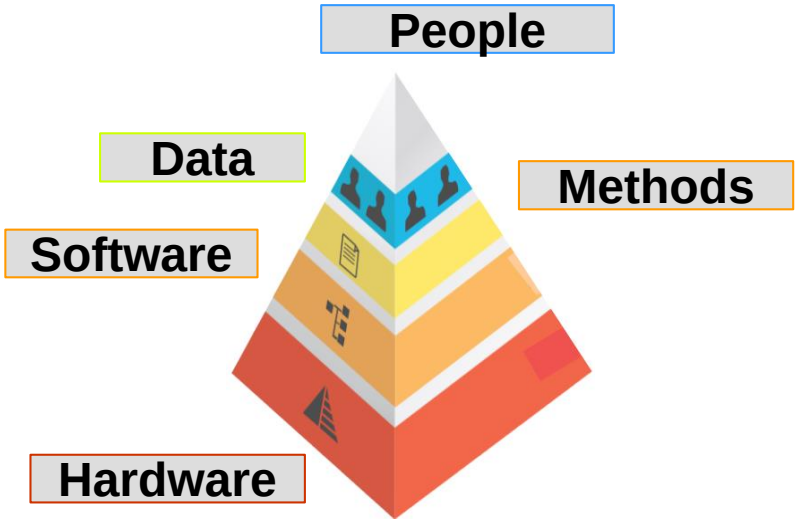
# PHIS

**PHIS:** Information System for phenomics based on **OpenSILEX**

**OpenSILEX: an open source software set**

- Methods, tools, components to implement information systems for experimental data in agriculture and environment

- Information System: Organized system for the collection, organisation, storage, exchange and treatment of information



# Data structuring

The **structuring of data** enables a computer system to perform store, retrieve, process data and Implement good practices:

- Make **FAIR data**
- **Flexible**
- Ability to allow **understanding (and reproduce) data processing**
- Ability to enforce DMP and Open Science

## Structuring of data based on 2 key elements:

- **Identification and Naming convention**
  - Objects: plants, plots, experiments, sensors, events, etc.
  - Persistent, unambiguous, resolvable, globally unique

- **Semantic and tagging (based on ontology set)**
  - Controlled vocabulary
  - Formalized relationships between entities
  - Data annotation and enrichment (search engine friendly)

# Data Structuring: PHIS approach

## PHIS → Ontology driven

**Scientific object instances** (plant, plant organ, plot, etc.) are formalized
Identified by **URI** standardized, unambiguous, shared, etc

**Events** (management, faults, meteo, etc) are formalized
Identified by **URI**

**Variables, Observations, Factors, Documents, Devices, Softwares** are associated with these Objects and Events
Identified by **URI**

**Organisation and linking of Objects and Events → done with a controlled semantic** (reference ontologies, vocabularies, thesaurus, taxonomies) and application Ontologies (RDF, OWL, SKOS)*

***Semantic Web Languages W3C**

# PHIS identification

**URI**

- **Standardized and** easy integration in Web application
- **Unambiguous**
- **Actionable** (resolvable,dereferencable)

URI → generated by tools under responsibility of local coordinator

**URI of plant**
*<http://phenome.fr/arch/2017/c17000118>*
**URI of pot:**
*<http://phenome.fr/arch/2013/pc13001542>*
**URI of cart:**
*<http://phenome.fr/arch/2013/ct1300123>*

**URI of cabin:**
*<http://phenome.fr/arch/2018/ac180015>*
**URI of camera:**
*<http://phenome.fr/arch/2018/ac180019>*

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

# PHIS identification

URL identifies what exists on the Web
URI identifies, on the Web, what exists
IRI  identifies, on the Web, in any language, what exists

**prefix m3p** : *<http://phenome.fr/arch>*

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

**URI of pot:**
*<m3p:2013/pc13001542>*

**URI of cart:**
*<m3p:2013/ct1300123>*

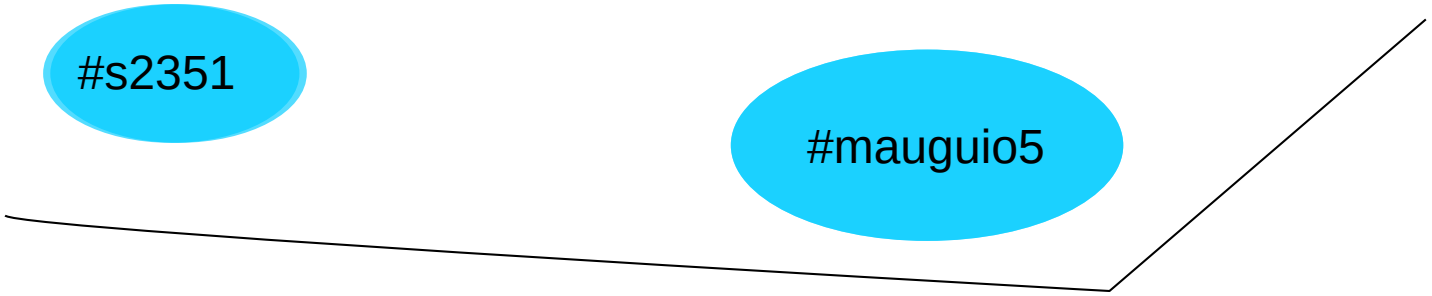
**URI of cabin:**
*<m3p:2018/ac180015>*

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

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

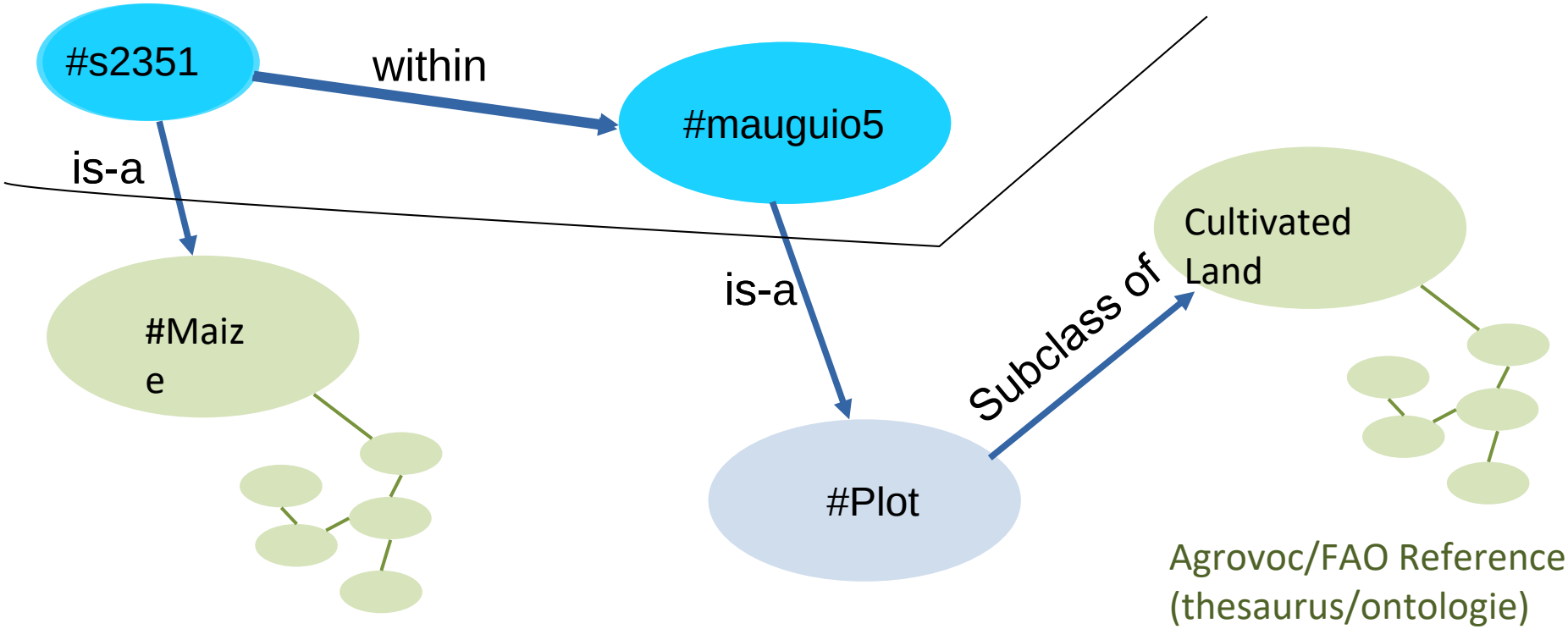
# PHIS and Semantics

**Metadata / ontologies provide the meaning of data**
**→ Link data elements using a controlled and shared vocabulary and also machine readable**

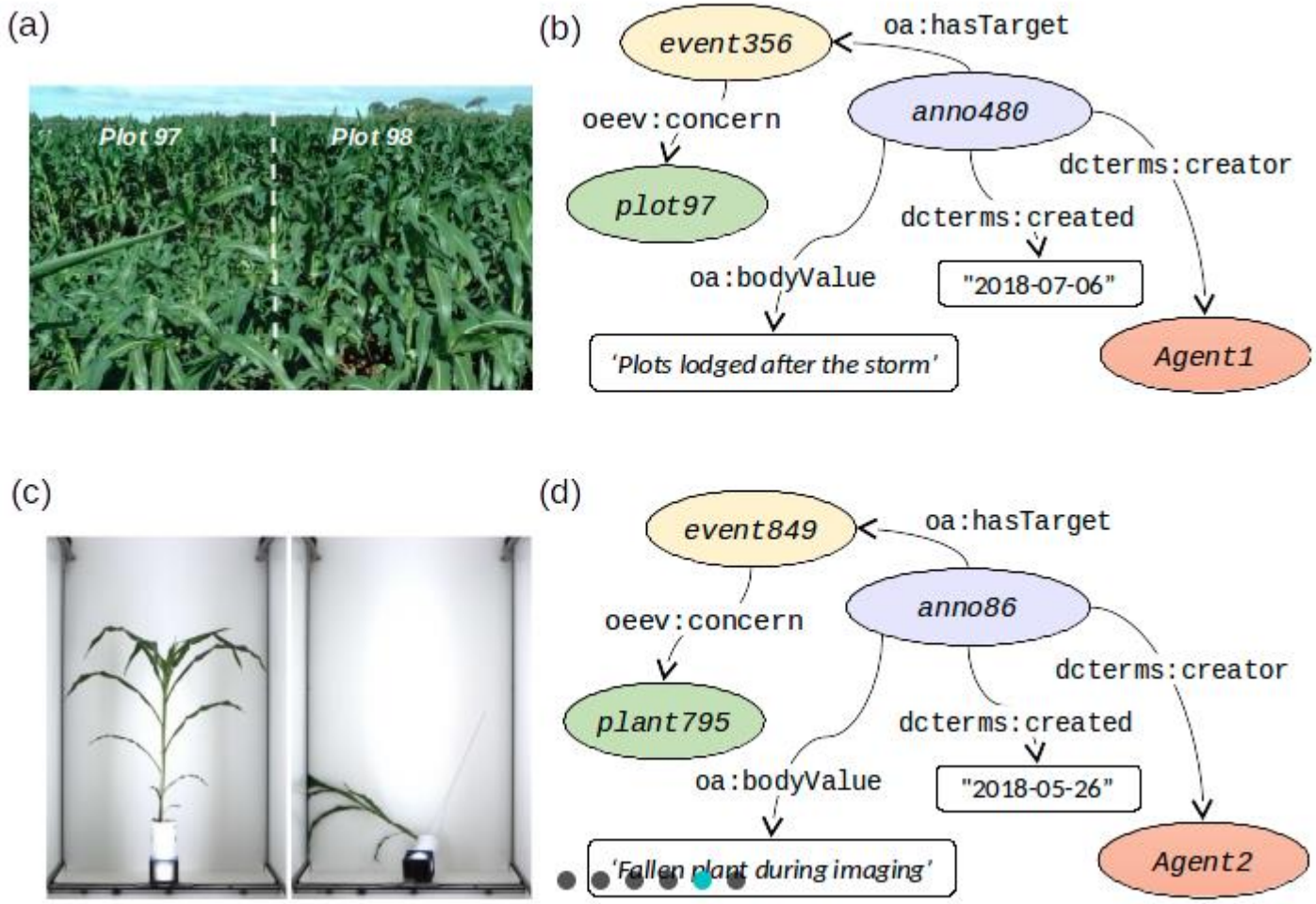


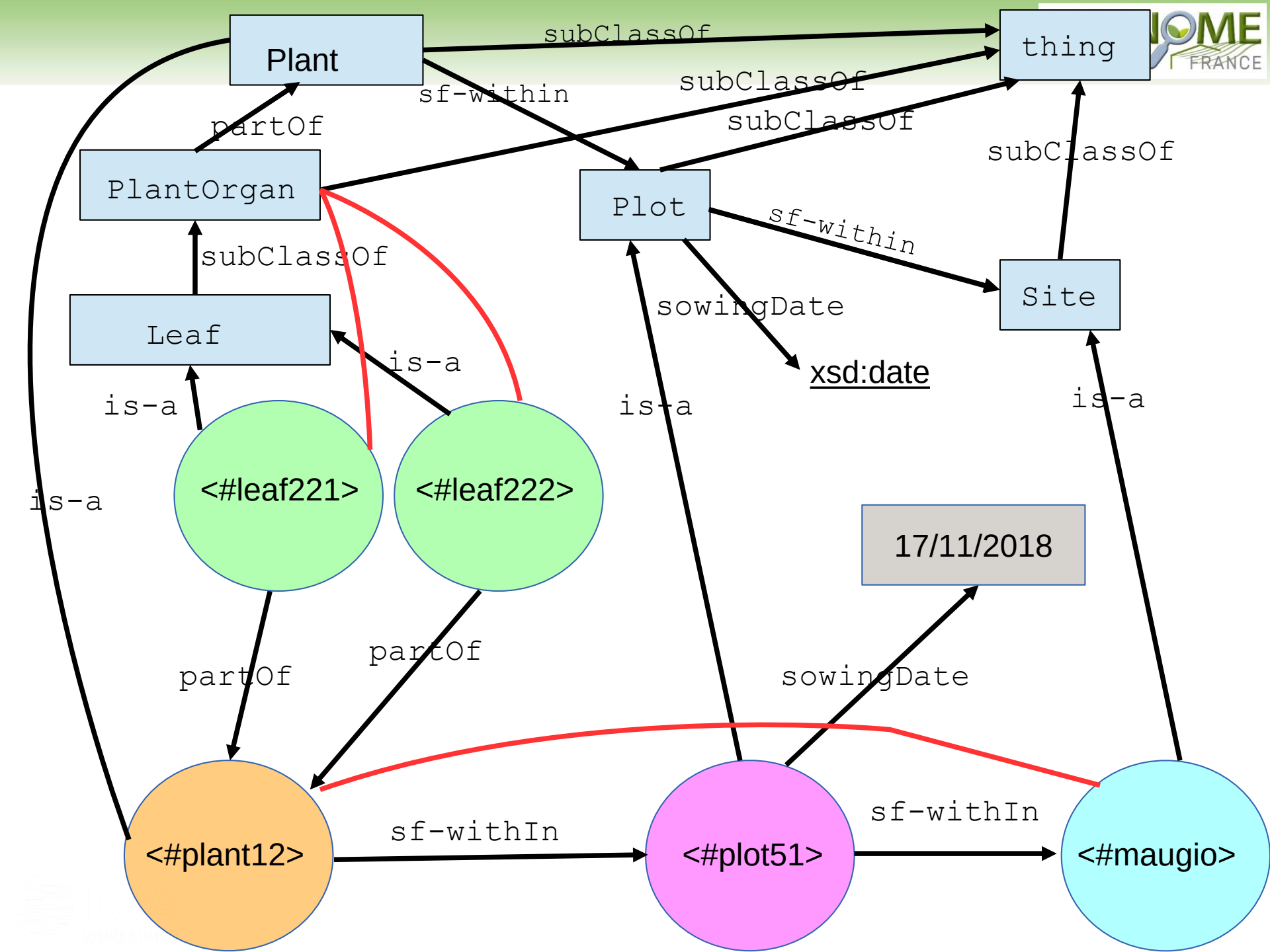
# PHIS and Semantics

**Metadata / ontologies provide the meaning of data**
**→ Link data elements using a controlled and shared vocabulary and also machine readable**
**→ Data are structured in a graph that can be queried**



# Ontology driven Information System




Plant
subClassOf
thing
sf-within
subClassOf
partOf
subClassOf
subClassOf
PlantOrgan
Plot
sf-within
subClassOf
Site
Leaf
sowingDate
xsd:date
is-a
is-a
is-a
is-a
<#leaf221>
<#leaf222>
is-a
17/11/2018
partOf
partOf
sowingDate
<#plant12>
sf-withIn
<#plot51>
sf-withIn
<#maugio>

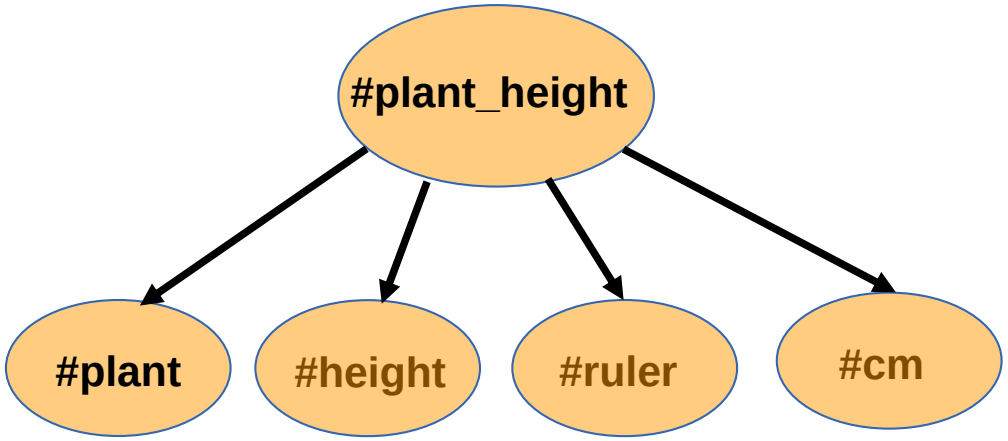
# Formalization of variables

What do we recommended in PHIS

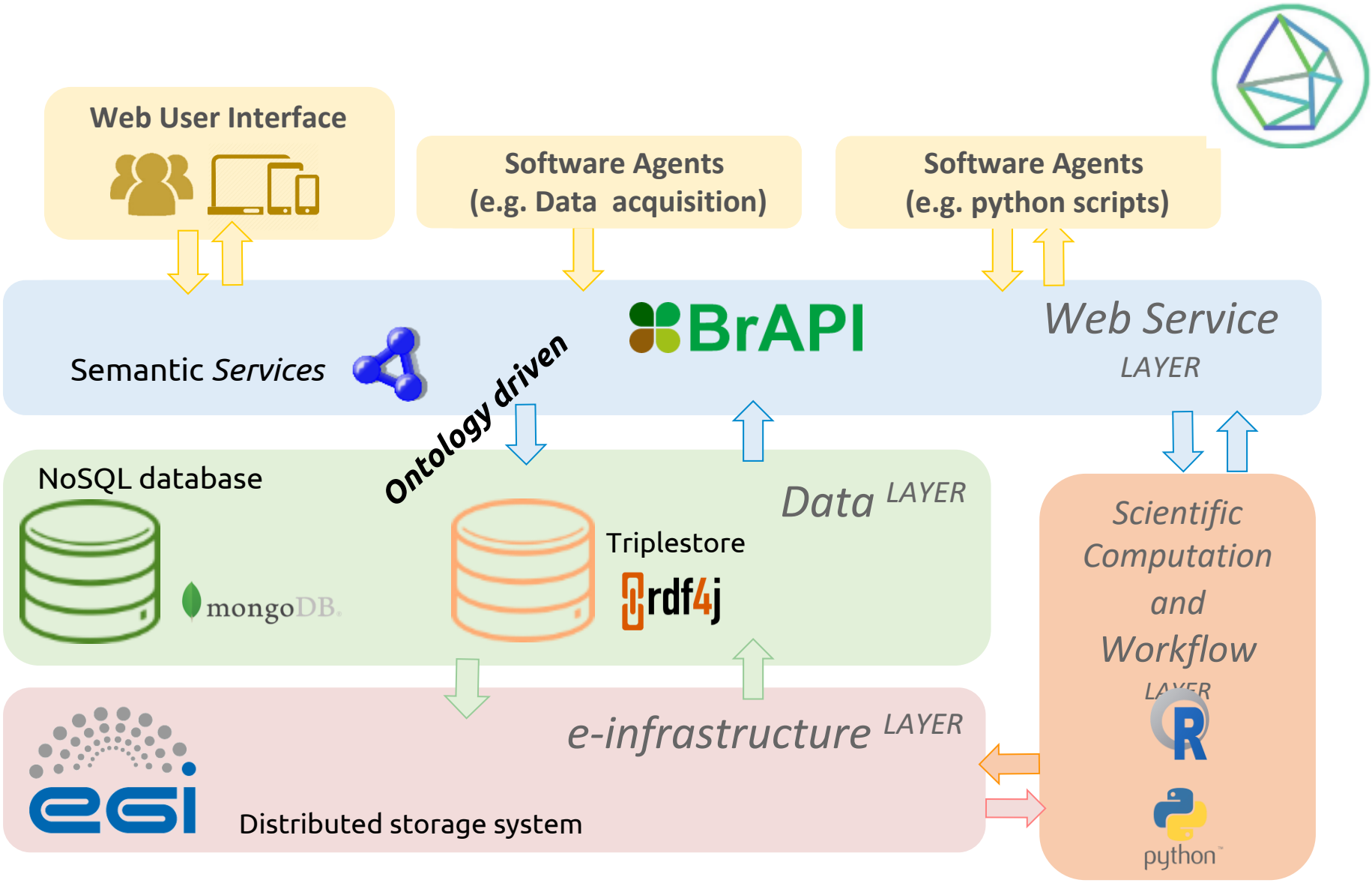
- Use URI (ID) for unambiguous variable (in global context)
- Actionable URI for an accessible description of variables
  Description can be **read by machine** and human
- Reuse existing variable description if available
- Use standardized representation schema for formalisation of new variable (and share it)

**Variable = Entity + Characteristic + Method + Scale**

**#plant_height = URI**



# OpenSILEX - PHIS Architecture



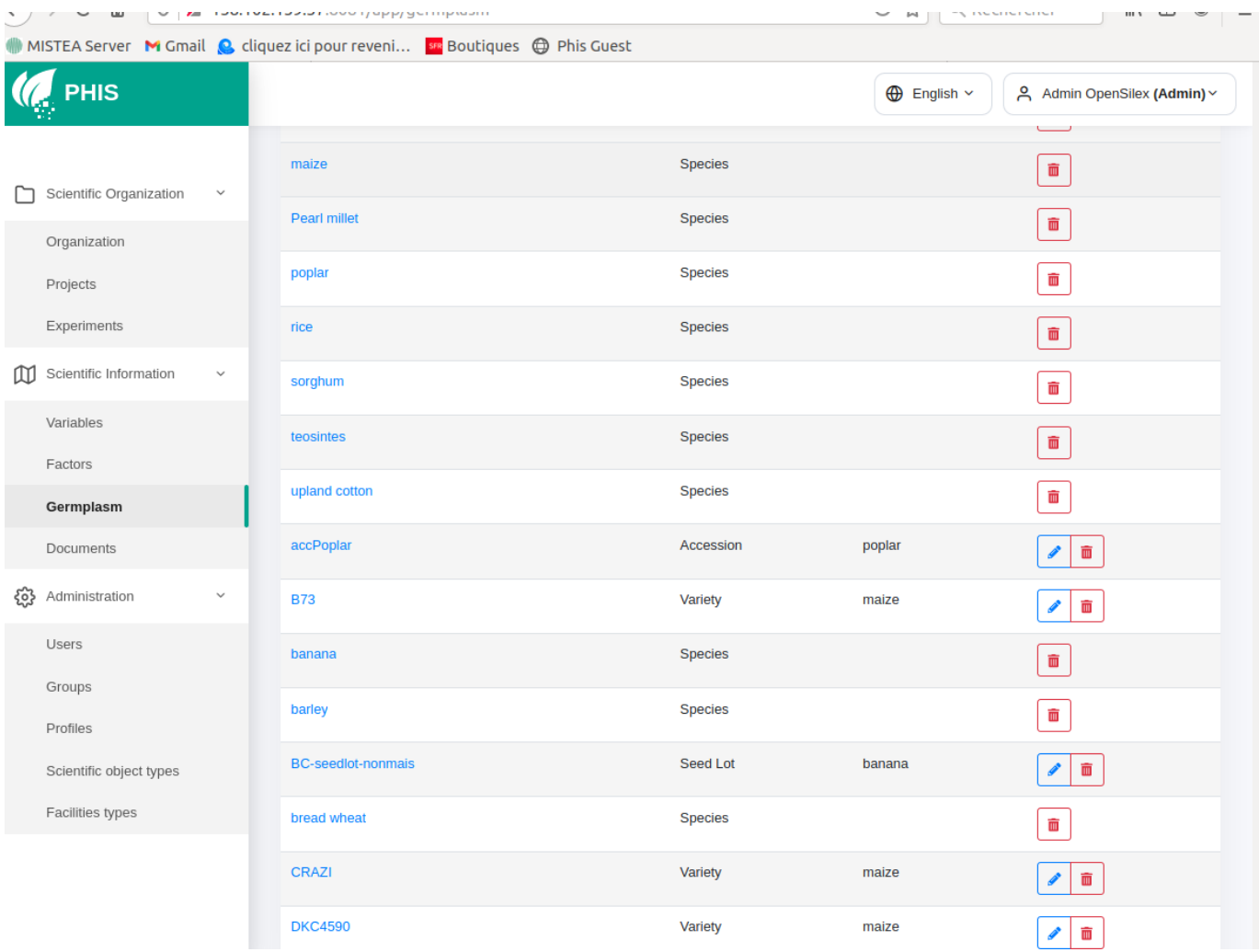
# PHIS User Interface

**PHIS Web User Interfaces** for the management of:

- Project information
- Experiment
- Facilities
- Devices

- Scientific objects
- Germplasms
- Experimental factors

- Data
- Data visualization
- Data provenance



# OpenSILEX / PHIS

- ✓**Allows management of huge and complex data**
- ✓**Enables and facilitates cloud computing (data center, EGI)**

  → **distributed computing, distributed storage, backup**
- ✓**Manage semantics (ontologies, standardized vocabularies)**
- ✓**Flexible design**
- ✓**Open technologies, Web APIs and portal interoperability**
- ✓**Provenance and reproducibility for data processing**
- ✓**Over 10 instances of PHIS** for various installations **(field and greenhouse)**
- ✓**Distributed PHIS instances: >1 petabytes of data**
- ✓**Other implementations of OpenSilex: WEIS, SunAGRI, Sixtine**
- ✓**Open Software** - support and development (MISTEA team)

# Conclusion

- ✓ A new generation of information systems (e.g. PHIS) is needed
- ✓ Giving value to complex data requires structuring according to FAIR principles
- ✓ A better formalization of concepts (using ontologies) and data is required for interdisciplinary research
- ✓ Advanced data management makes data available for AI and data analytics
- ✓ Support and training available

# PHIS and OpenSILEX

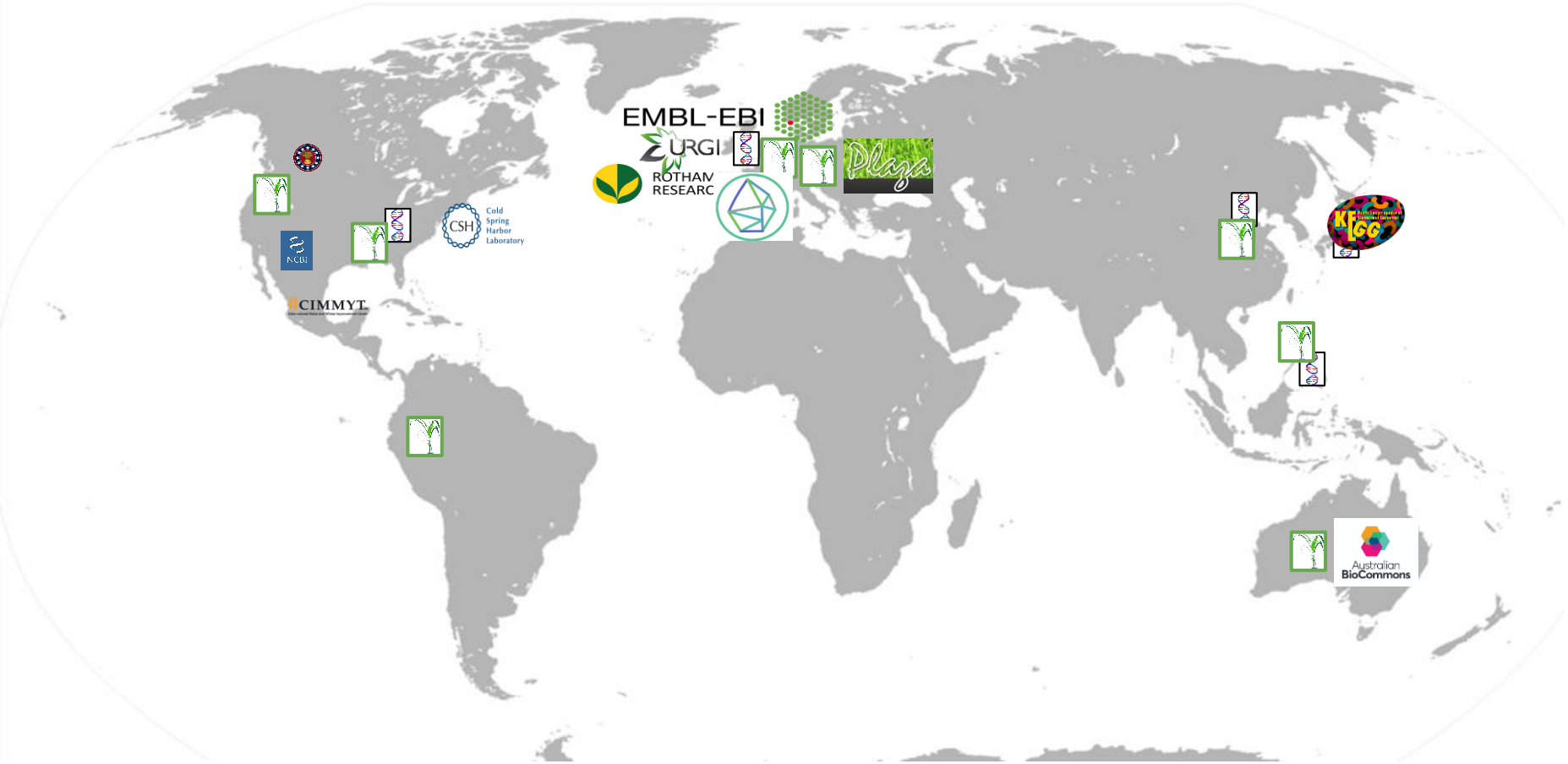
- **PHIS** demonstration
  - http://phis.inra.fr/
  - Research paper: https://nph.onlinelibrary.wiley.com/doi/full/10.1111/nph.15385

- How to contribute to OpenSILEX?
  - Github repository: https://github.com/OpenSILEX/
  - Developer documentation: https://opensilex.github.io/docs-community-dev/

- User documentation of the version in development:
  - https://opensilex.github.io/phis-docs-community/

# Global Information systems

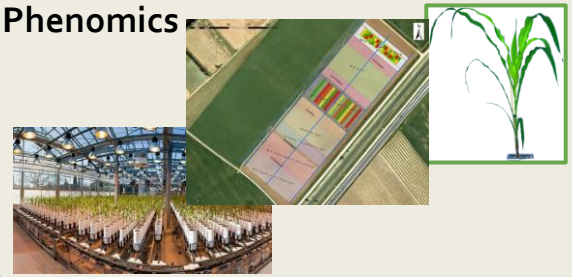
Dispersed data  Heterogenous data  Dedicated repositories & Archives



# Data standardisation and exchange

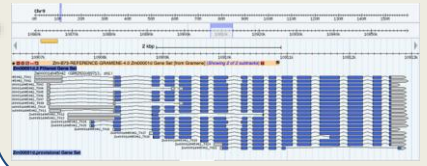
Environment / Phenome / Genomic / *omic / Genetic

**Phenomics**



Findable Accessible Interoperable Reusable

**Genetics**
**Genomics**
**Omics**



**Plant Breeding**
Genetic variations by Traits

**Climate Change Studie**
Genotype by Environment

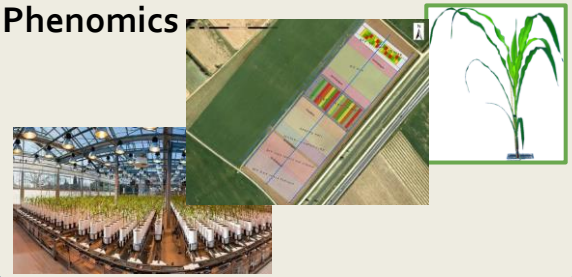
Findable Accessible Interoperable Reusable

**Environment**

# PLANT use case

Environment / Phenome / Genomic / *omic / Genetic

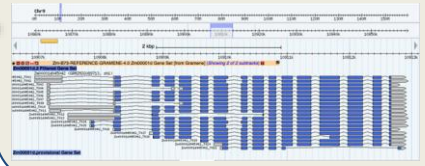
**Phenomics**



Dispersed
Heterogenous
Getting Standardized

Findable Accessible Interoperable Reusable

**Genetics**
**Genomics**
**Omics**



Mostly centralized
Homogenous data
Heterogenous metadata

**Plant Breeding**
Genetic variations by Traits

**Climate Change Studie**
Genotype by Environment

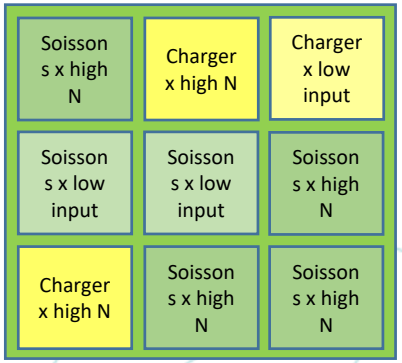
Findable Accessible Interoperable Reusable

Dispersed
Heterogenous

**Environment**

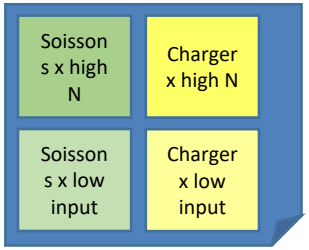
# Phenotyping data life cycle

« Raw » data, pheno/env measures, variables

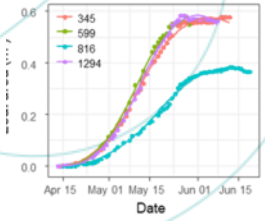



Derivation, Reduction

« computed » data, reduced, indicators




| Genotype | Treatment | N input | Date | Rep | Fusariose |
|---|---|---|---|---|---|
| Soissons | low input | 15,32253129 | 15/11/2011 | 1 | 5 |
| Soissons | low input | 15,31430556 | 16/11/2011 | 2 | 7 |




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

| | | | | |
|---|---|---|---|---|
| 661300270 | Ardon | 2004 | 45.645632645603683 | 12/01/2004 284.3 |
| 661300270 | Ardon | 2005 | | |
| 661300444 | Ardon | 2004 | 38.96112577281653 | 12/01/2004 228.8 |
| 661300444 | Ardon | 2005 | | |
| 661300312 | Cavallermaggiore | 2004 | 52.4 | 01/01/2004 249.9 |
| 661300312 | Cavallermaggiore | 2005 | | |
| 661300371 | Cavallermaggiore | 2004 | 45.74 | 01/01/2004 230.2 |
| 661300371 | Cavallermaggiore | 2005 | | |
| 661300487 | Cavallermaggiore | 2004 | 72.52 | 01/01/2004 309.8 |
| 661300487 | Cavallermaggiore | 2005 | | |
| 661300585 | Cavallermaggiore | 2004 | 71.739999999999995 | 01/01/2004 305.7 |
| 661300585 | Cavallermaggiore | 2005 | | |
| 661300468 | Headley | 2004 | 45.27 | 01/01/2004 |
| 661300468 | Headley | 2005 | | |
| 661300469 | Headley | 2004 | 70.930000000000007 | 01/01/2004 |
| 661300469 | Headley | 2005 | | |
| 661300533 | Headley | 2004 | 57.67 | 01/01/2004 258.8 |

# Plant Phenotyping Life cycle

## Raw data long term conservation

**Data acquisition**

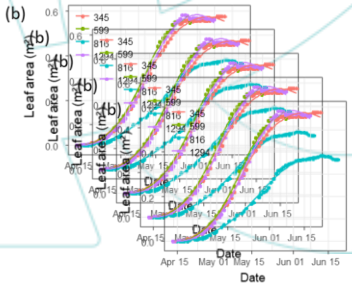
- **VARIABLES**
- Plant/microplot level
- Traceability
- Raw measures
- Data Cleaning
- Platform IS (Emphasis IS, PHIS, …)
- Analysis Reproducibility
- Provenance

**Data computation**

- **INDICATORS**
- Statistical integration
- Genotype level (mostly)
- New computation for each scientific question
- One raw dataset  many computed datasets

**Data publication**

- One Data Publication by datasets.
- **Platform IS**
  - Phenomic, plant level
- **FAIR Data Repositories**
  - Reduced




| Genotype | traitement | Fusariose |
|---|---|---|
| Soisson | low input | 5 |
| Soisson | high N | 7 |
| Charger | low input | 1 |
| Charger | high N | 2 |

Variety charger is resistant to fusariose under intensiv cultural practice

# Plant Phenotyping Life cycle

## Raw data long term conservation

**Data acquisition**

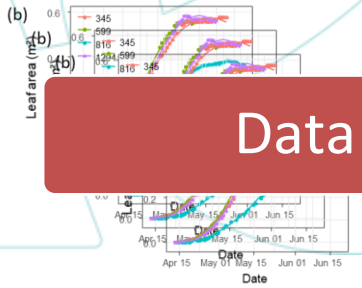
- **VARIABLES**
- Plant/microplot level
- Traceability
- Raw measures
- Data Cleaning
- Platform IS (Emphasis IS, PHIS, …)
- Analysis Reproducibility
- Provenance

**Data computation**

- **INDICATORS**
- Statistical integration
- Genotype level (mostly)
- New computation for each scientific question
- One raw dataset → many computed datasets

**Data publication**

- One Data Publication by datasets.
- **Platform IS**
  - Phenomic, plant level
- **FAIR Data Repositories**
  - Reduced




**Data**

| Genotype | traitement | Fusariose |
|---|---|---|
| Soisson | low input | 5 |
| Soisson | | 7 |
| Charger | low input | 1 |
| Charger | high N | 2 |

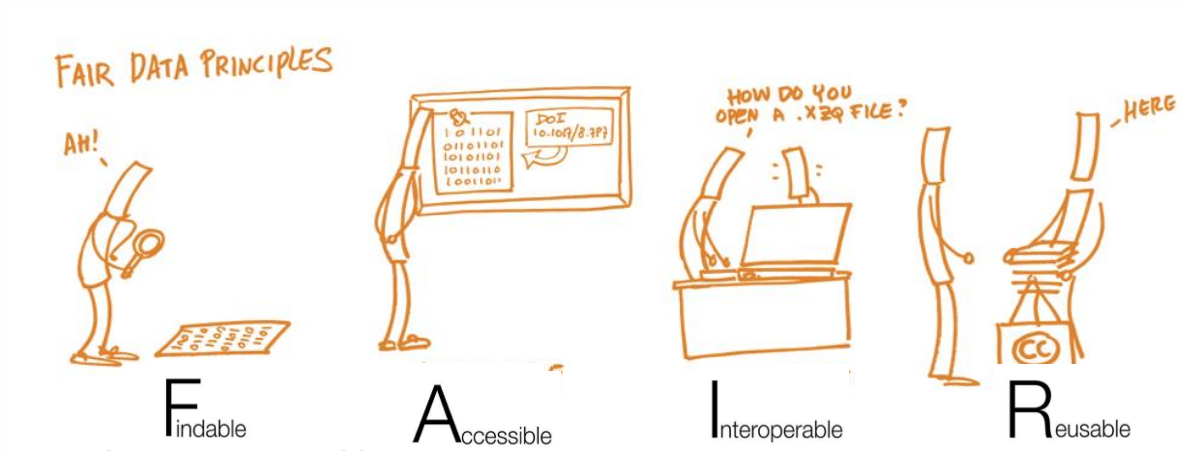
**Knowledge**

Variety charger … to … nder … cultural practice

# Open science through FAIR data principles

Sustainable data access over decades

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




**Ids**
**Index**
**Metadata**
**Description**

**Open Protocole**
**Perenial**
**Metadata**

**Semantics**
**Linked Data**
**Vocabularies**

**License**
**Well described**
**Provenance** (origin, process, methodology)
**Standards**

**Discovery** **Repositories** **Standards**

# FAIR data repositories

https://projet-recherchedatagv.ouvrirlascience.fr/

RÉPUBLIQUE FRANÇAISE
Liberté
Égalité
Fraternité

**Découvrir Recherche Data Gouv**

Accueil | Pourquoi Recherche Data Gouv ? | Le projet | Les acteurs du projet | Ateliers de la donnée | Actualités


À propos de Recherche Data Gouv

INRAE la science pour la vie, l'humain, la terre | UGA Université Grenoble Alpes | Université de Lorraine | cnrs | Université de Lille | Université de Strasbourg | Université de Paris | Université Paris Nanterre

# FAIR data repositories

https://projet-recherchedatagv.ouvrirlascience.fr/

https://data.inrae.fr/

Portail Data INRAE > INRAE > Omics Dataverse > URGI Plant and Fungi Dataverse >

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

Version 2.0

Millet, Emilie J.; Pommier, Cyril; Buy, Mélanie; Nagel, Axel; Kruijer, Willem; Welz-Bolduan, Therese; Lopez, Jeremy; Richard, Cécile; Racz, Ferenc; Tanzi, Franco; Spitkot, Tamas; Canè, Maria-Angela; Negro, Sandra S.; Coupel-Ledru, Aude; Nicolas, Stéphane D.; Palaffre, Carine; Bauland, Cyril; Praud, Sébastien; Ranc, Nicolas; Presterl, Thomas; Bedo, Zoltan; Tuberosa, Roberto; Usadel, Björn; Charcosset, Alain; van Eeuwijk, Fred A.; Draye, Xavier; Tardieu, François; Welcker, Claude, 2019, "A multi-site experiment in a network of European fields for assessing the maize yield response to environmental scenarios", https://doi.org/10.15454/IASSTN, Portail Data INRAE, V2, UNF:6:zF9w0A2f+MHeW7maeeXJWA== [fileUNF]

Modalités d'accès au dataset

Contact | Partager

Statistiques d'utilisation sur les datasets

Citer le dataset

Pour en apprendre davantage sur le sujet, consulter le document Data Citation Standards [en].

**Description**

This dataset comes from the European Union project DROPS (DR
panel of 256 maize hybrids was grown with two water regimes (irr
2012 and 2013, respectively, spread along a climatic transect fror
one site in Chile in 2013. This resulted in 29 experiments defined

a scenario but largely differ between scenarios.

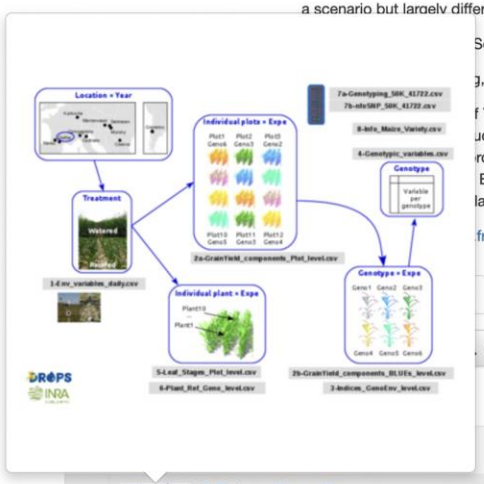
Sciences; Agricultural Sciences

, Genotype-by-environment interaction, Multi-envi

f Yield in Europe: Allelic Effects Vary with Drought a
ude Welcker, Willem Kruijer, Sandra Negro, Aude C
rde, Cyril Bauland, Sebastien Praud, Nicolas Ranc
Bedo, Xavier Draye, Björn Usadel, Alain Charcoss
lant Physiology, 172 (2) 749-764 doi: 10.1104/pp.1

fr/ephesis/ephesis/viewer.do#dataResults/trialSetI

Chercher




0_Data_outline.pdf

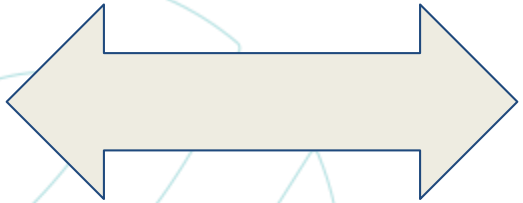
application/pdf - 1.0 Mo - 27 mars 2019 - 412 téléchargements

MD5: 1a11142041623c0364c90acae645ae78

This file contains the outline of the files organization and content. It describes the different measurement scales available: Weather and soil water data at daily time step (1). Phenotypic data at the plot level in each experiment (2a) and at the genotypic level in each experiments (2b and 3). Variable calculated at the genotype level (4). Phenotypic data at the plant level in each experiment for the reference genotype (5 and 6). Genotyping data available for each genotype (7a and 7b). Description of the

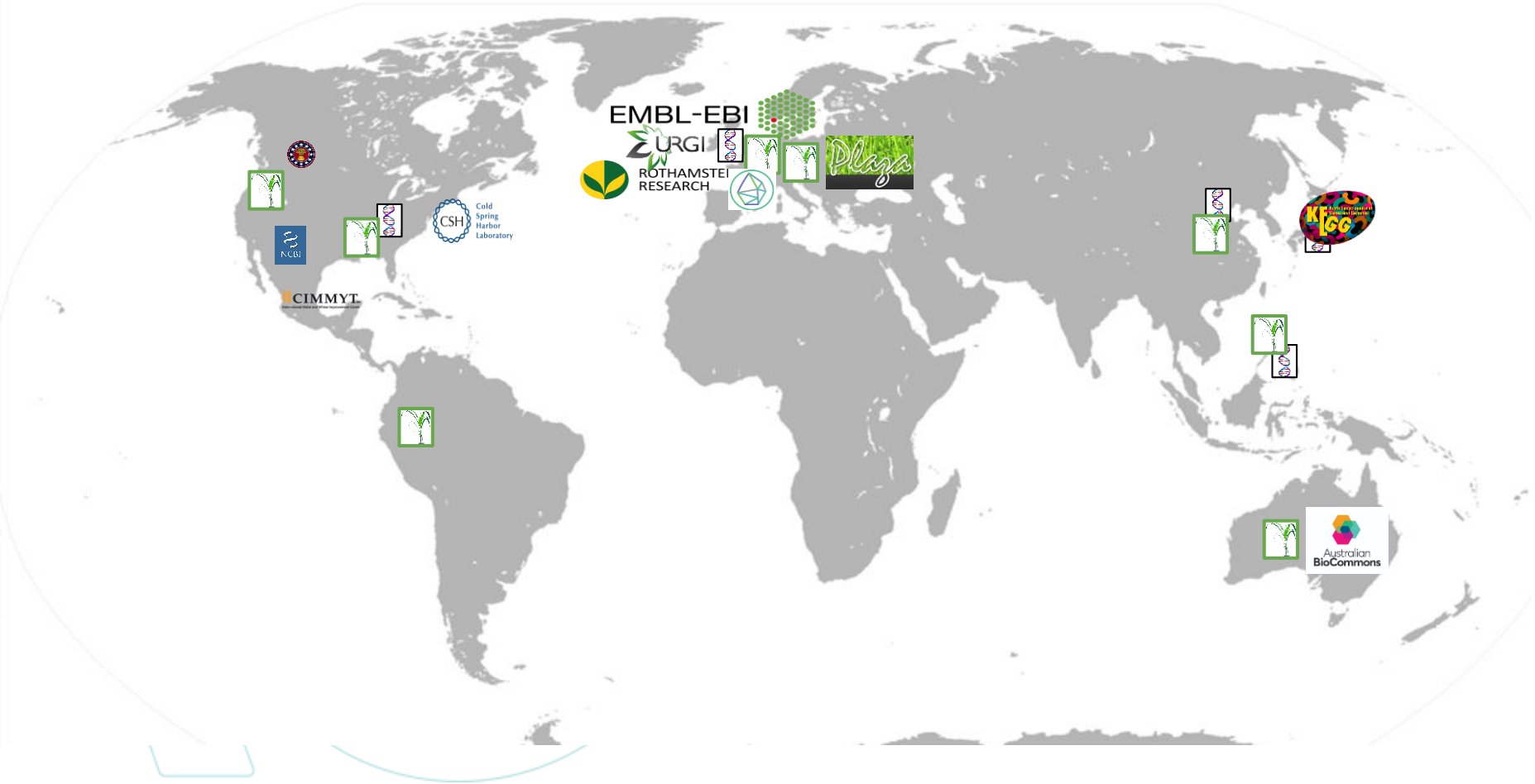
https://www.phenome-e

# FAIR data repositories



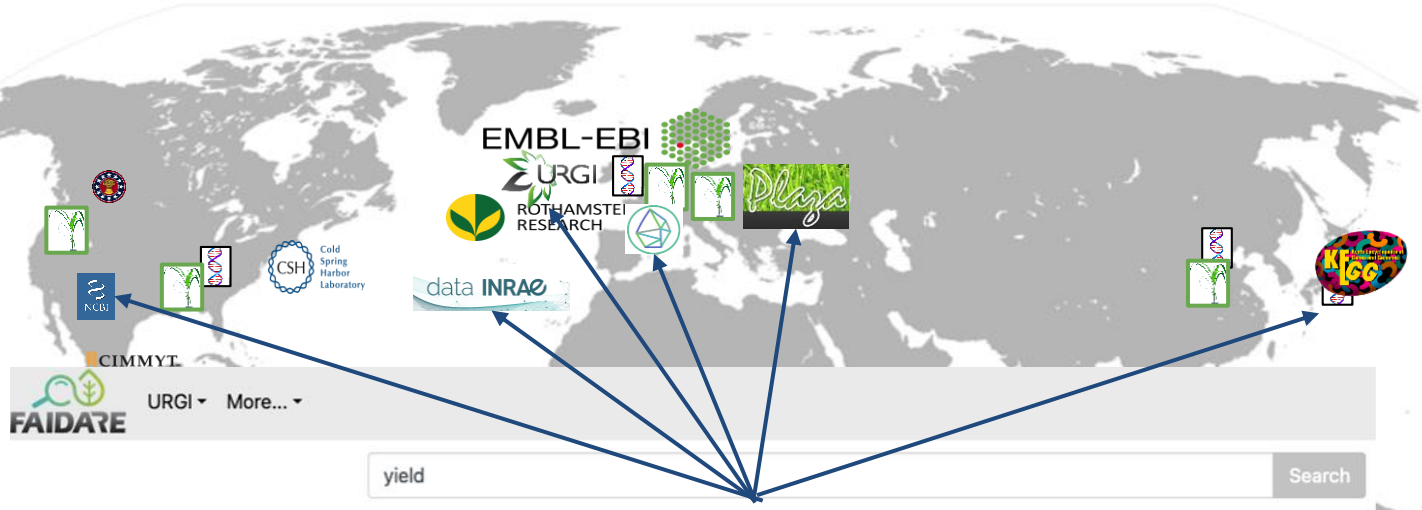
# Global Data discovery portal

Dispersed data  Heterogenous data  Dedicated repositories & Archives



# Global Data discovery portal

Dispersed data  Heterogenous data  Dedicated repositories & Archives




**Species** *(21)*

Filter on Species...

**Data type**

- ☐ Bibliography [151]
- ☐ None [5]

**Ontology annotation** *(30)*

Filter on Ontology annotat

Expand search

**Database**

- ☐ OpenMinTeD@GnpIS [151]
- ☐ GEBIS [5]

Results 1 to 20 of 156

*10.3389/fpls.2018.00529* - *OpenMinTeD@GnpIS*

Bibliography Triticum Triticum aestivum

Global QTL Analysis Identifies Genomic Regions on Chromosomes 4A and 4B Harboring Stable Loci for **Yield**-Related Traits Across Different Environments in Wheat (Triticum aestivum L.). 2018 Global QTL Analysis Identifies Genomic Regions on Chromosomes 4A and ... (expand)

*10.1186/s12864-019-6005-6* - *OpenMinTeD@GnpIS*

Bibliography Triticum Triticum aestivum

Genome-wide association study reveals new loci for **yield**-related traits in Sichuan wheat germplasm under stripe rust stress. 2019 Genome-wide association study reveals new loci for **yield**-related traits in Sichuan wheat germplasm under stripe rust stress Ba ... (expand)

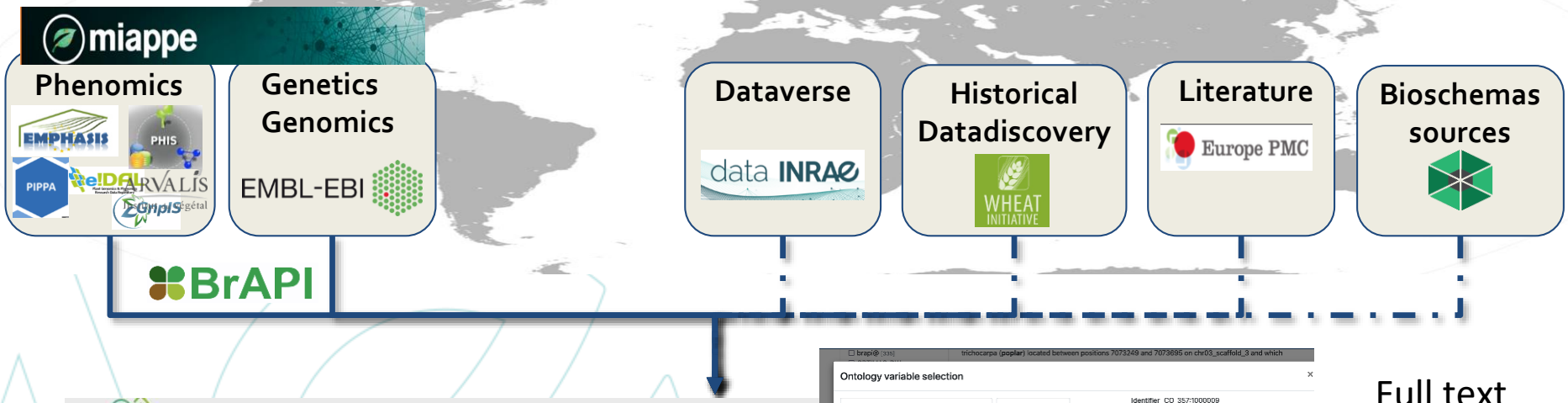
*10.1186/s12863-019-0785-1* - *OpenMinTeD@GnpIS*

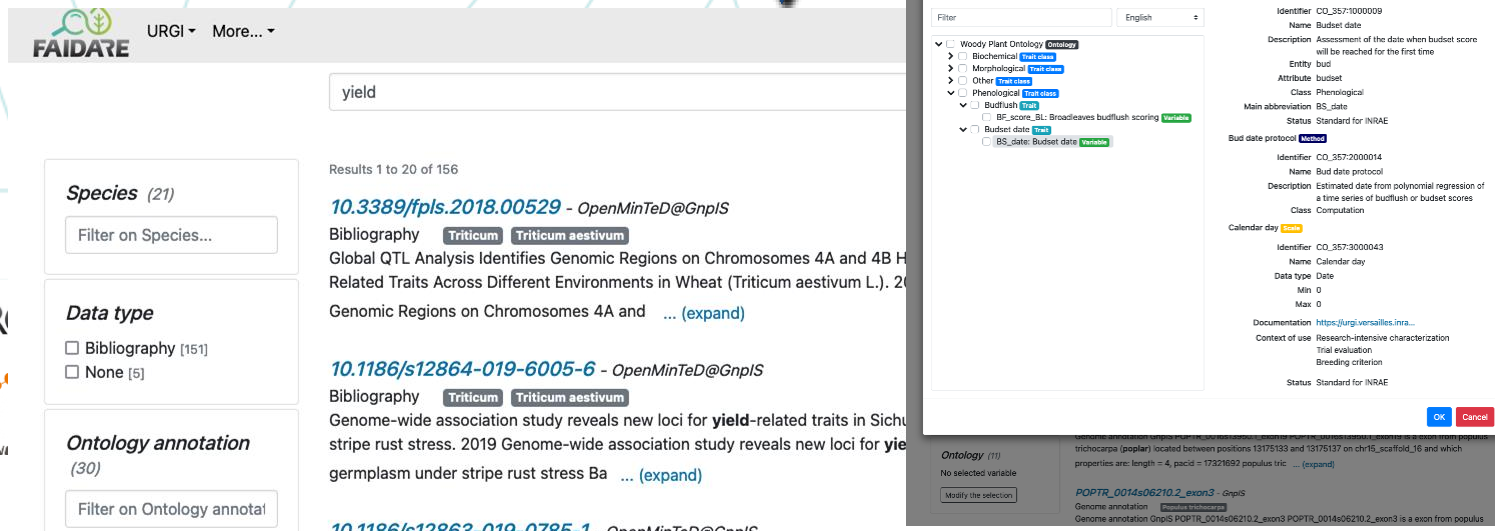
Bibliography Triticum

Accuracy of genomic selection for grain **yield** and agronomic traits in soft red winter wheat. 2019 Accuracy of genomic selection for grain **yield** and agronomic traits in soft red winter wheat Background Genomic selection has the potential to increase genetic ... (expand)

# FAIDARE: Global Data discovery portal

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




Full text
+
Fine criteria
+
**Link back**

# Data standards for FAIR

## Semantic

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

*Biologist driven*

## Structure

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

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

## Technical

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

# Phenotype Structure Standard

Minimal Information About Plant Phenotyping Experiment

www.miappe.org


Many stakeholders

Elixir, Emphasis, Bioversity, North American PPN

Open Community:

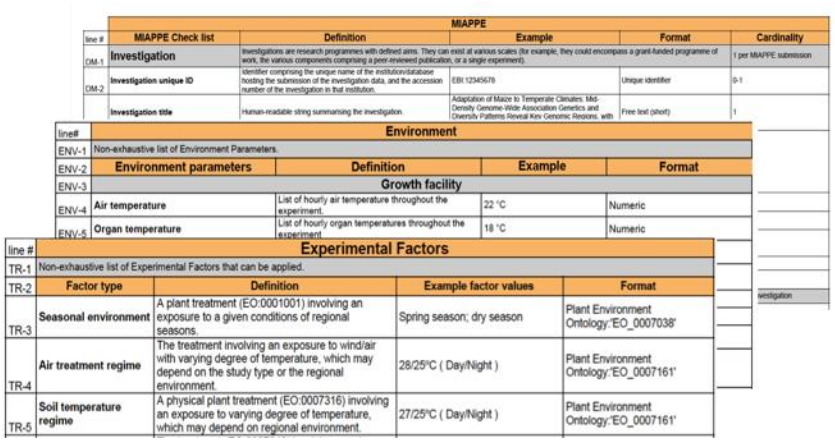
Request for comments

Github Feature requests

Mailing lists

Meetings & Workgroups

Crops and woody plants



| | MIAPPE | | | | |
|---|---|---|---|---|---|
| line # | MIAPPE Check list | Definition | Example | Format | Cardinality |
| DM-1 | Investigation | Investigations are research programmes with defined aims. They can exist at various scales (for example, they could encompass a grant-funded programme of work, the various components comprising a peer-reviewed publication, or a single experiment). | | | 1 per MIAPPE submission |
| DM-2 | Investigation unique ID | Identifier comprising the unique name of the institution/database hosting the submission of the investigation data, and the accession number of the investigation in that institution. | EBI:12345678 | Unique identifier | 0-1 |
| | Investigation title | Human-readable string summarising the investigation. | Adaptation of Maize to Temperate Climates: Mid-Density Genome-Wide Association Genetics and Diversity Patterns Reveal Key Genomic Regions, with | Free text (short) | 1 |

| line# | Environment | | | |
|---|---|---|---|---|
| ENV-1 | Non-exhaustive list of Environment Parameters. | | | |
| ENV-2 | Environment parameters | Definition | Example | Format |
| ENV-3 | Growth facility | | | |
| ENV-4 | Air temperature | List of hourly air temperature throughout the experiment. | 22 °C | Numeric |
| ENV-5 | Organ temperature | List of hourly organ temperatures throughout the experiment. | 18 °C | Numeric |

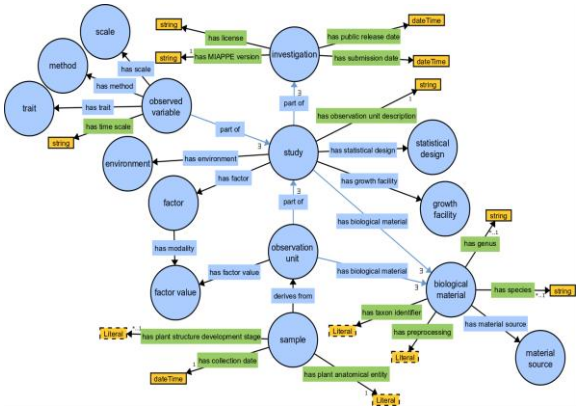
| line # | Experimental Factors | | | |
|---|---|---|---|---|
| TR-1 | Non-exhaustive list of Experimental Factors that can be applied. | | | |
| TR-2 | Factor type | Definition | Example factor values | Format |
| TR-3 | Seasonal environment | A plant treatment (EO:0001001) involving an exposure to a given conditions of regional seasons. | Spring season; dry season | Plant Environment Ontology:'EO_0007038' |
| TR-4 | Air treatment regime | The treatment involving an exposure to wind/air with varying degree of temperature, which may depend on the study type or the regional environment. | 28/25°C ( Day/Night ) | Plant Environment Ontology:'EO_0007161' |
| TR-5 | Soil temperature regime | A physical plant treatment (EO:0007316) involving an exposure to varying degree of temperature, which may depend on regional environment. | 27/25°C ( Day/Night ) | Plant Environment Ontology:'EO_0007161' |

Papoutsoglou *et al*. (2020) Enabling reusability and interoperability of plant phenomic datasets with MIAPPE 1.1. New Phytol, 227:260-273; **https://doi.org/10.1111/nph.16544**

# Phenotype Technical Standard, MIAPPE Implementations

Ontology, OWL Implementation

- https://github.com/MIAPPE/MIAPPE-ontology
- http://agroportal.lirmm.fr/ontologies/PPEO
- Data model representation
- Formal concepts and constraints

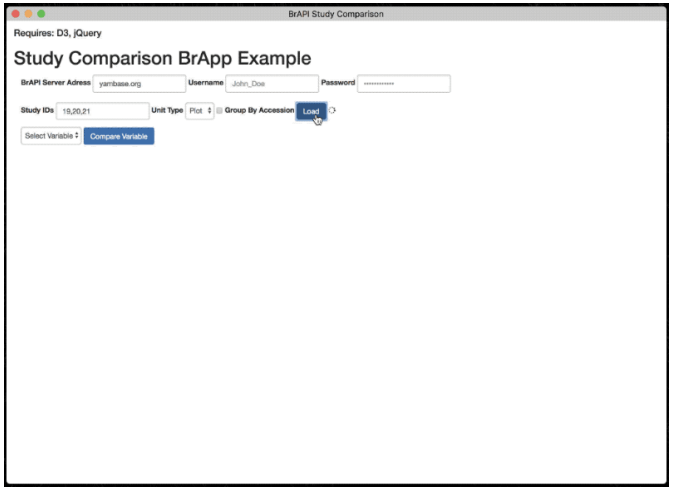



File Archive

- ISA Tab: data + metadata
- RO Crate studies

Web Services

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




Adoption

- Phenospex
- ELIXIR
- EMPHASIS
- European Genbanks (EURISCO)

# Take Home Message

Research Data Reuse, including by yourself, Public or not

- → Complex
- → Relies on FAIR data principles and in particular
  - Data Management
    - Information systems
    - Data Management Plans
  - Data repositories


  - Data standards
  - Data search portals

- Support and Guidances
  https://rdmkit.elixir-europe.org/

# Aknowledgments


**Elixir Plant community & platforms**
Beier S., Gruden C., Pommier C., Coppens F, Scholz U., Lange M., Contreras B., Adam Blondon AF, Faria D, Chavez I, Miguel C, Droedsbek B, Finkers R, Papoutsoglou E, Olster R, Ramsak Z, …

**Emphasis**
Tardieu F, Usadel B, Alic I, Arend D, Junker A, Poorter H, Neveu P, Pierushka R, Shur U…
And many more!

**Crop Ontology**
Arnaud E, Laporte MA, …

**MIAPPE community**
ELIXIR Plant Community,
Krajewsky P, Cwiek H, Tardieu F, Usadel B, Arend D, Arnaud E, Junker A, King G, Laporte MA, Poorter H, Reif J, Rocca-Serra P, Sansone SA, Kersey P,
And many more!

**Breeding API**
Selby P, Mueller L, Robbins K, Backlund JE, … ,
And many more!

**H2020 AGENT**
N. Stein (IPK, coord), P. Kersey (RBGK), M. Alaux (INRAE), S. Weise (IPK), C. Pommier (INRAE), M. Lange (IPK), R. Finkers (WUR), *J. Destin (INRAE)*


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