Apport des ressources Semantiques pour la gestion de données

Contact: Pascal.Neveu@inra.fr

Data Challenge

Context: more and more data!

- Cheap storage capacity and high speed network e.g. **1 Gigabyte price** : $400K in 1980, $10K in 1990, $10 in 2000, **now less than $0.01**
- Many heterogeneous devices, simulations, machine learning, Internet data sources (Open, collaborative, etc) are available

Make data valuable!

- **Knowledge discovery**
- **Decision support**
- **New services and Open Science**

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

Need: A new generation of Information Systems

INRA
SCIENCE & IMPACT

Complex Data

Different scales

Genome

Organ

Plant

Field

Region

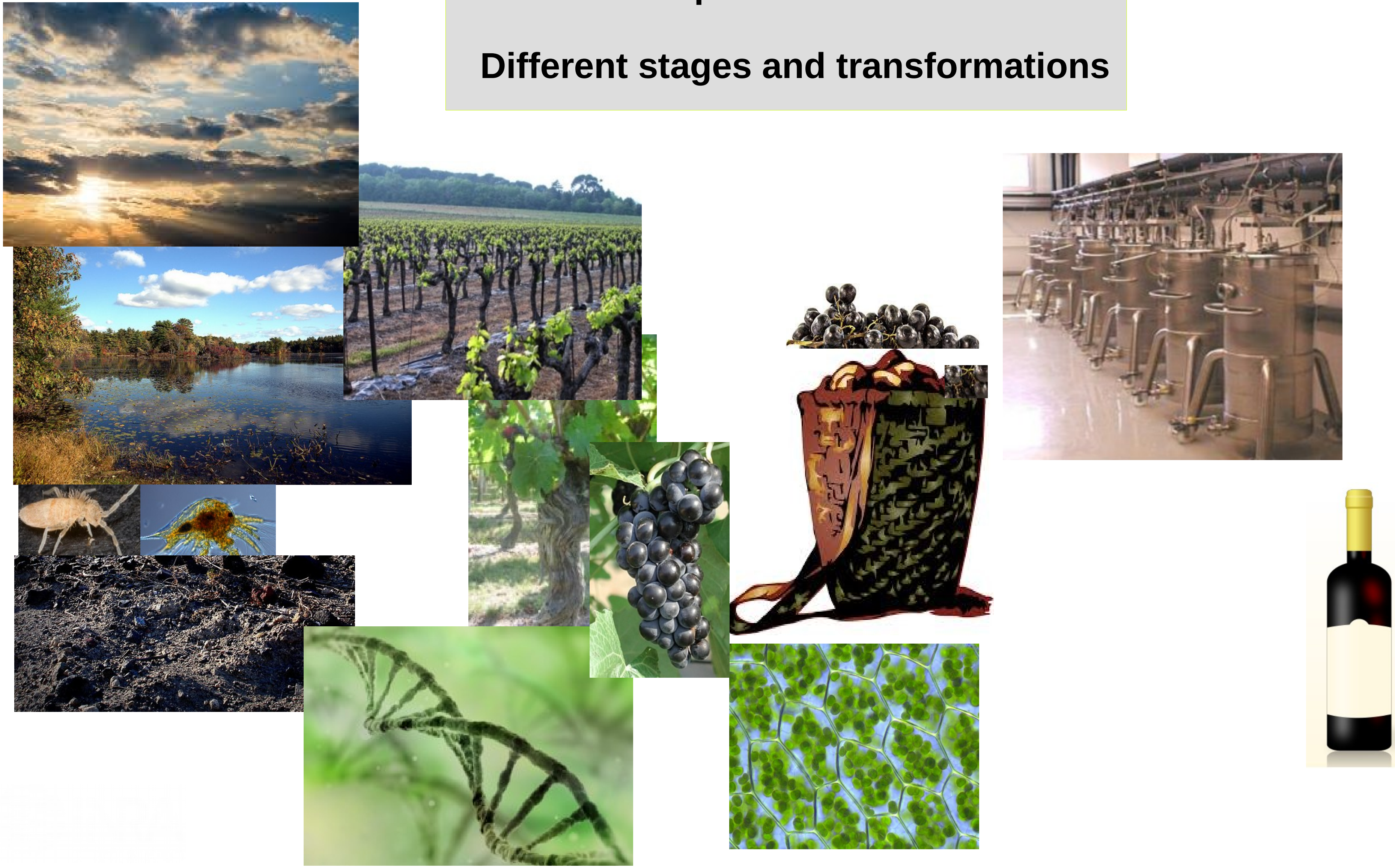
Complex Data

Different interactions

Complex Data

Different crops

Different stages and transformations



Complex Data

From various contexts

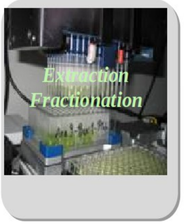
« omics » Platforms

Various data complex types

Genomics

Composition and the structure of biopolymers

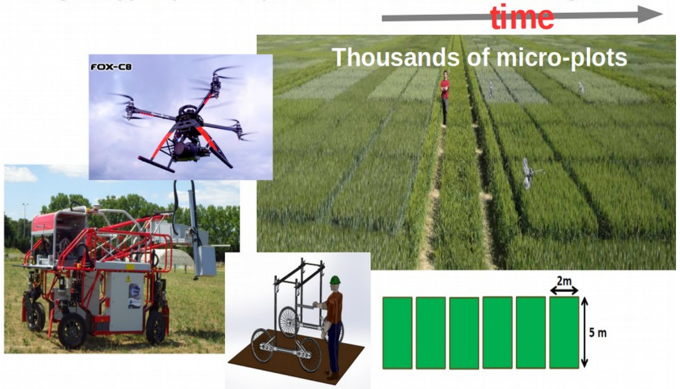
Quantification of metabolites and enzyme activities



Field Platforms

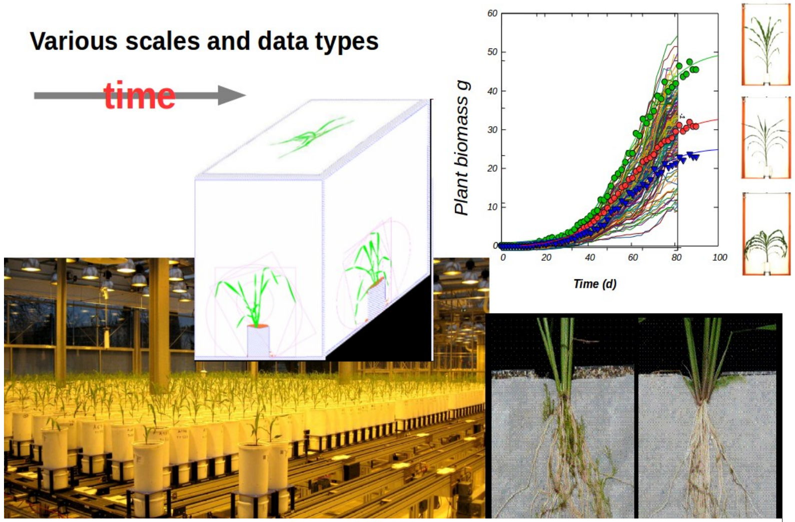
Various scales and data types

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



Green house Platforms

Various scales and data types

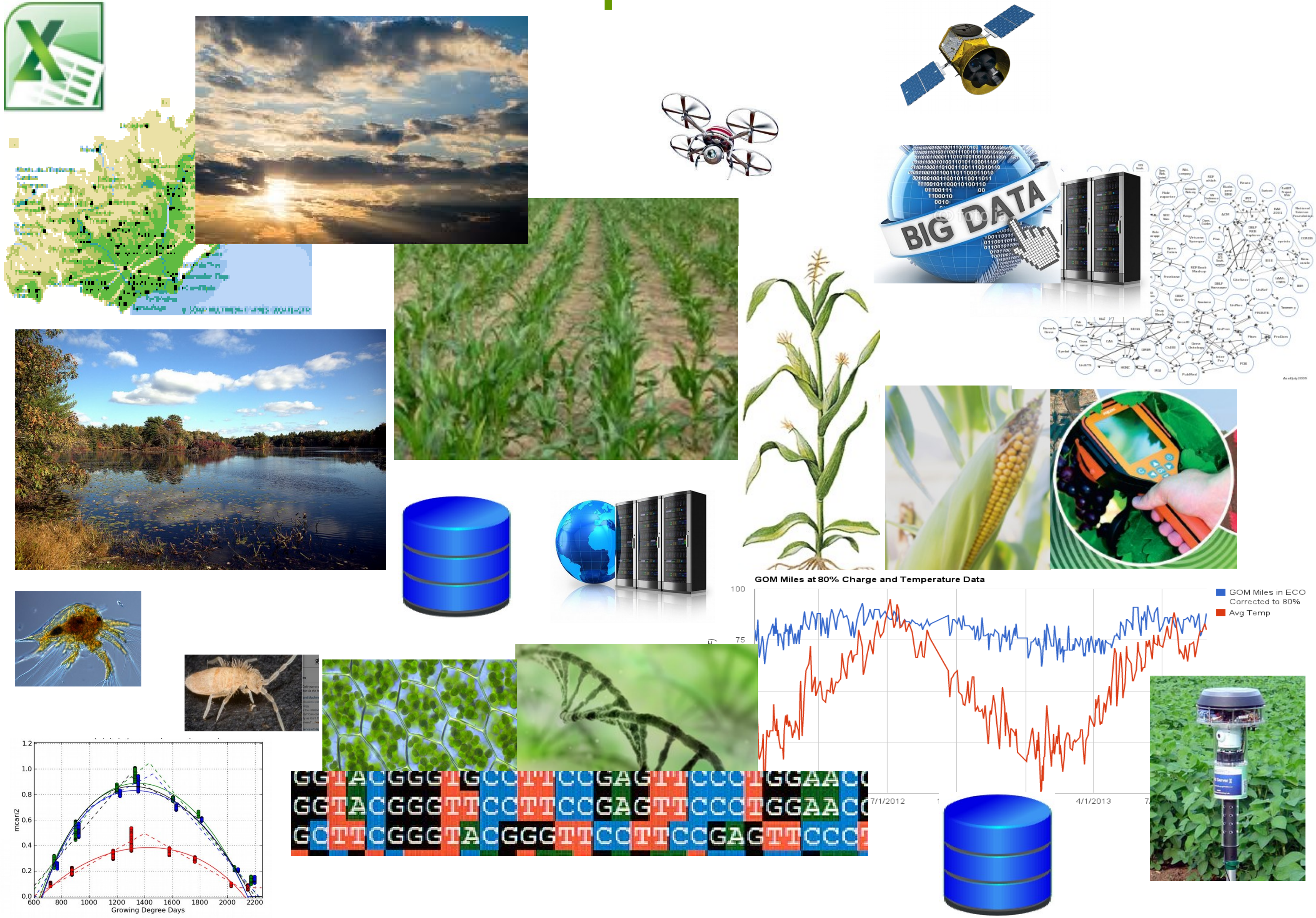


Farm Platforms

Various scales and data types from thousands of farms

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

Complex Data


BIG DATA
GOM Miles at 80% Charge and Temperature Data
100
75
7/1/2012
4/1/2013
GOM Miles in ECO Corrected to 80%
Avg Temp
GGTACGGGTGCCTTCCGAGTTCCCTGGAACC
GGTACGGGTTCCTTCCGAGTTCCCTGGAACC
GCTTCGGGTACGGGTTCCTTCCGAGTTCCCT
1.2
1.0
0.8
0.6
0.4
0.2
0.0
mcari2
600
800
1000
1200
1400
1600
1800
2000
2200
Growing Degree Days

Complex Data

- **Orphan data → Worthless!**
- **Data has value if they are grouped...**
  (re)analyse, meta-analyse, visualisation, etc

Context

Experiments or Observations

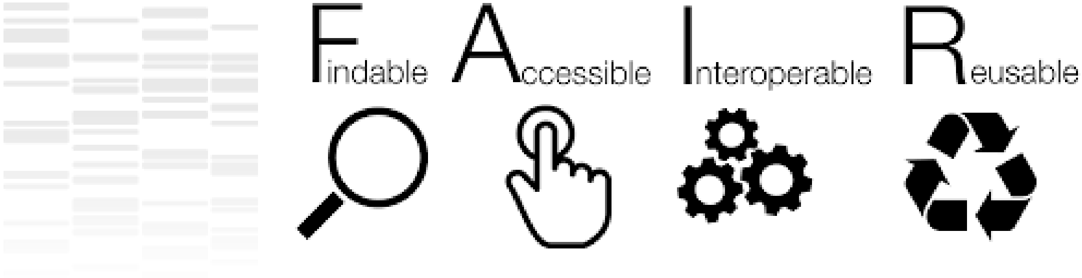
- Expensive, require a lot of resources and often very hard
- Cannot be reproduced

Scientific projets generate large and complex datasets

Strong needs of transparence and reproducibility

But re-analyses meta-analyses and new analyses

→ **impossible without information (metadata)**



Findable: **PID** (globally unique), indexed in portals, standardized and relevant metadata

Accessible: open and standardized protocols (internet protocols), **licence rights**

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

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

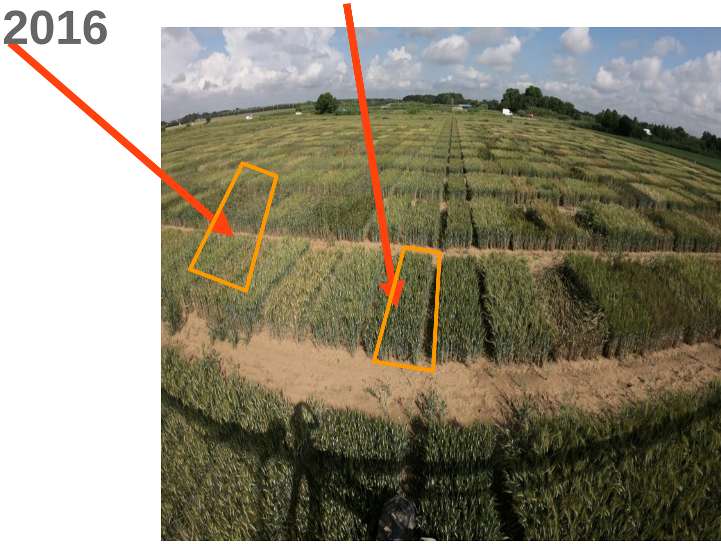
Some common mistakes we do

- Metadata in file names (not standardized, very often not machine readable, reduces metadata quantity)

  → 2017-Paris-Syrah-irrig-goblet.csv

  → 2017-St-Paul-Merlot-guyot-not_irrig.csv

- Same name for several variables
- Several names for same variable
- Sharing unstable variables
- Data are stored on personal computer
- Sofware parameters (calibration, etc) are lost
- Ambiguous ID
- X-Y position are lost
- Faults are not described
- no or few data links
- etc



Plot566 in 2016

Plot566 in 2017

Complex Data

How to structure data ?

Structuration

Data structure 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

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 ontologie set)**

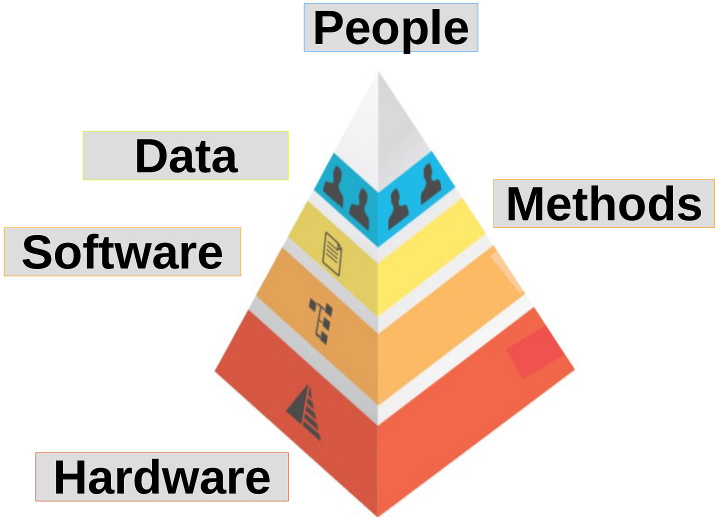
- Controlled vocabulary
- Formalized relationships between entities
- Data annotation and enrichment

OpenSILEX

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



Structuration : PHIS approach

PHIS is the Information System for Phenomics based on OpenSILEX

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

- Identified by **URI** standardized, unambiguous, shared, etc

Events (management, faults, meteo, etc)

- Identified by **URI**

Variables, Documents, Observations, Software 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**

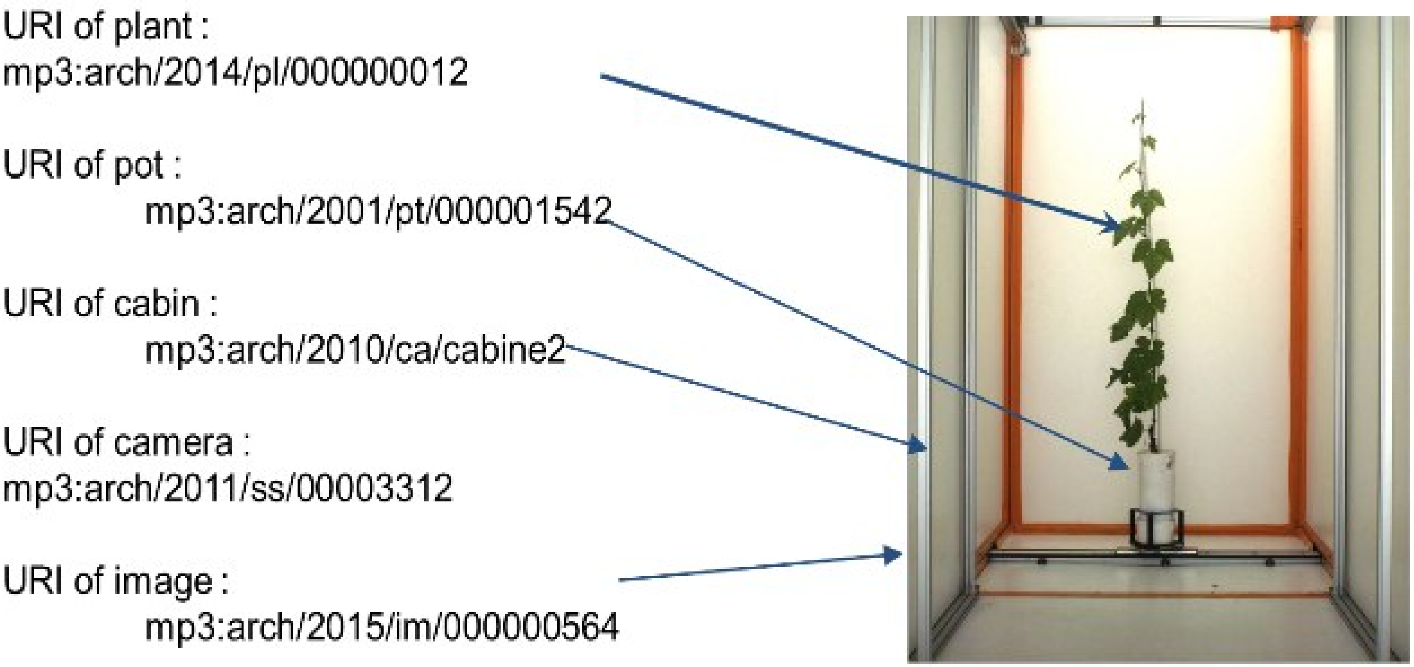
PHIS Identification

URI string used to identify a resource (Web standardized syntax)

→ **Standardized (easy integration in Web application)**

`http://subdomain.yourdomain.topdomain/path/identifier`

Possibility to use prefix: m3p: <http://lepse.inra.fr./>



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

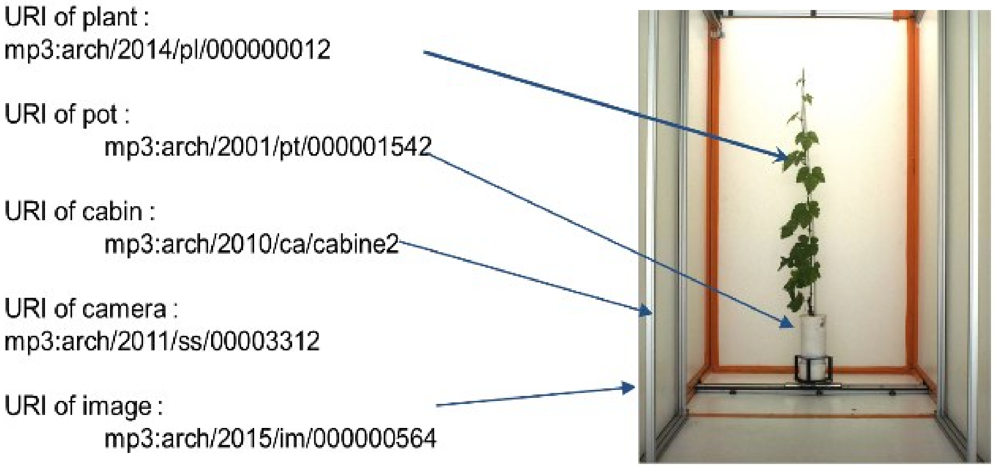
URI of image :
mp3:arch/2015/im/000000564

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.



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

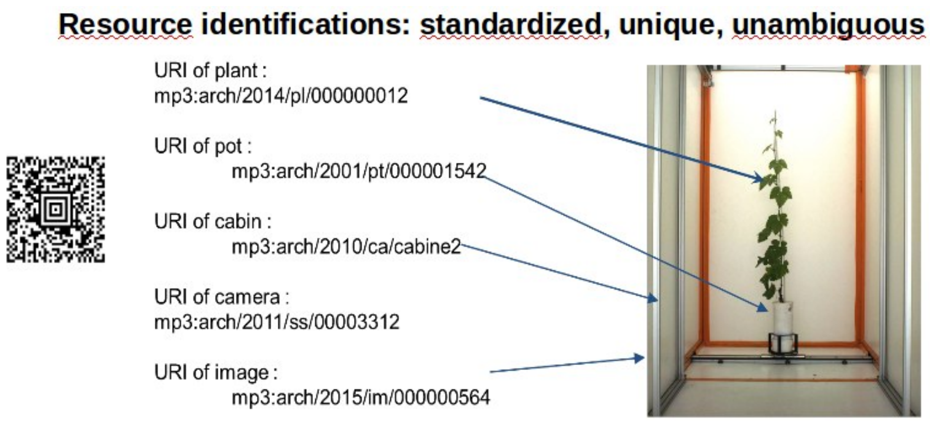
URI of image :
mp3:arch/2015/im/000000564

PHIS Identification

URI string used to identify a resource (Web standardized syntax)

- **Standardized**
- **Unambiguous, globally unique**
- **Resolvable (actionable, dereferencable)**
- **Persistent (services: B2HANDLE, PIC, PURL, etc)**

URI → the scientific responsible and generated by tools



Resource identifications: standardized, unique, unambiguous

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

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

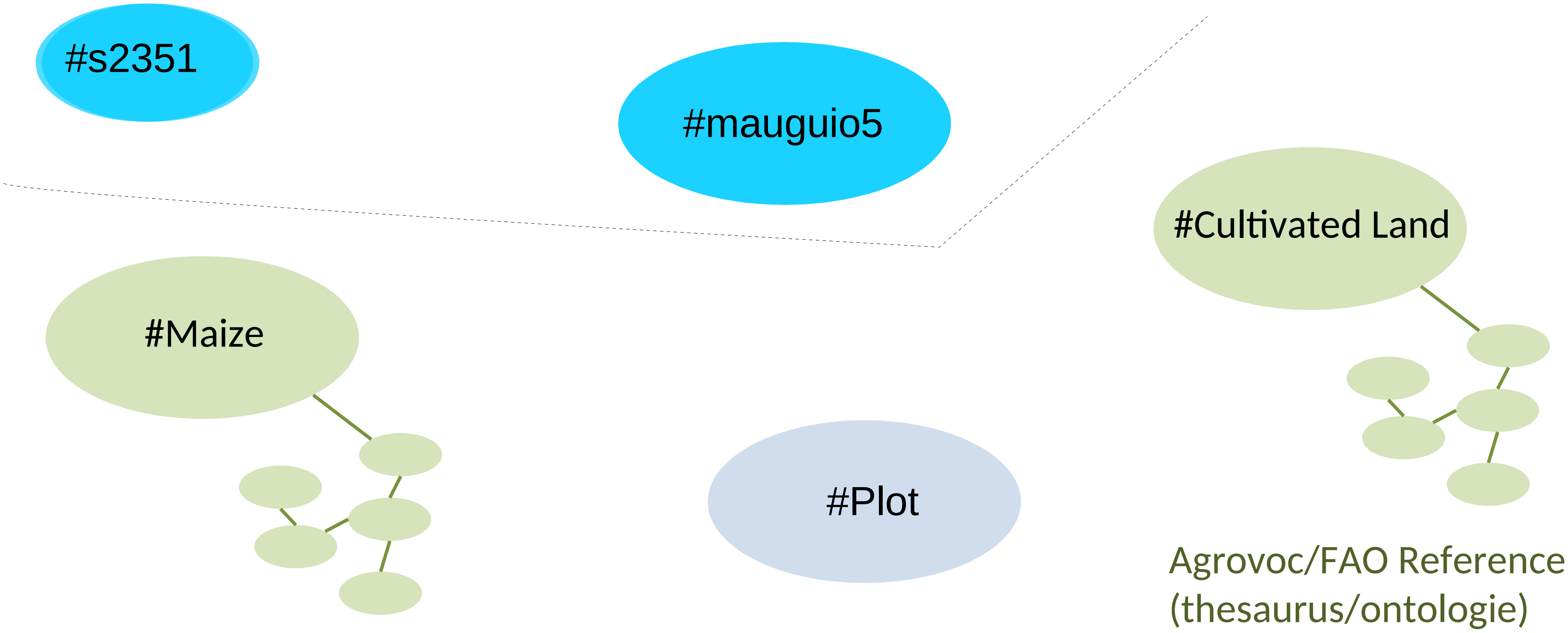
URI of cabin :
mp3:arch/2010/ca/cabine2

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

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

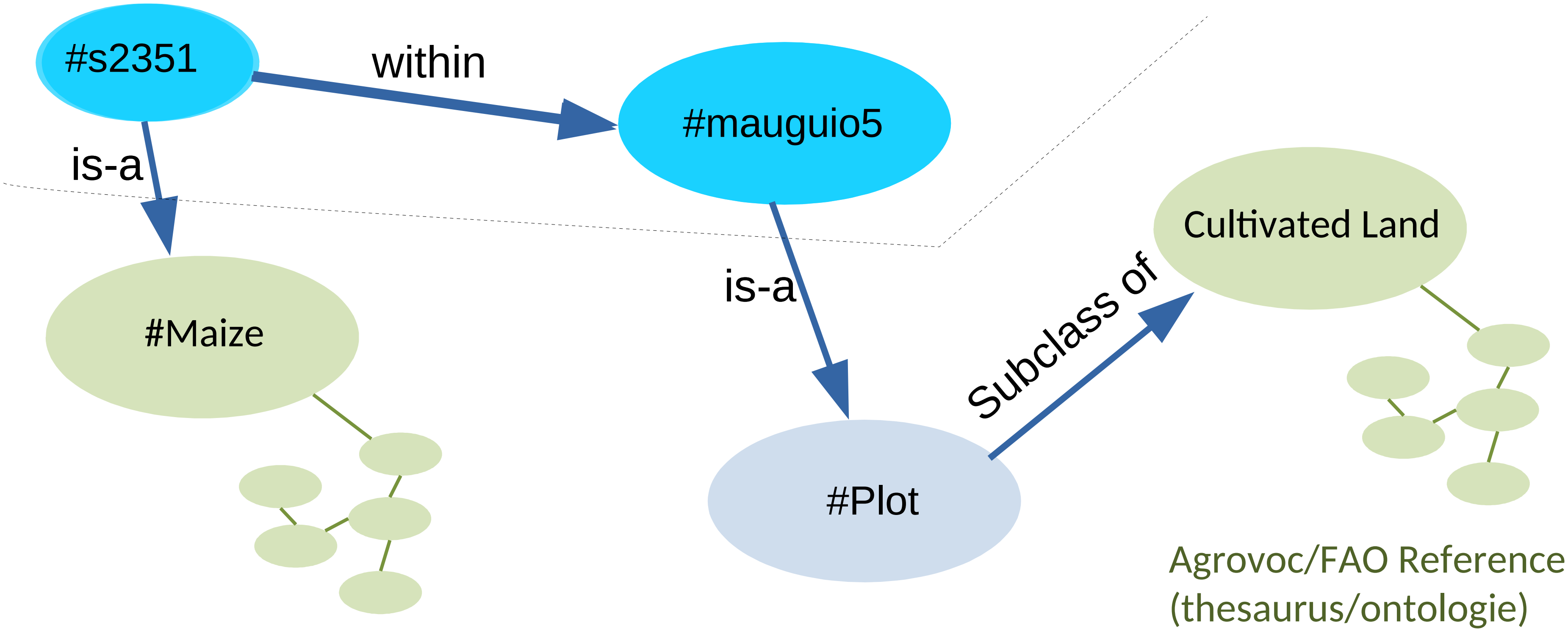
PHIS

- Metadata / ontologies provide the meaning of data
  → Link each data element to a controlled, shared, vocabulary and machine readable
  → Structure the data in a graph

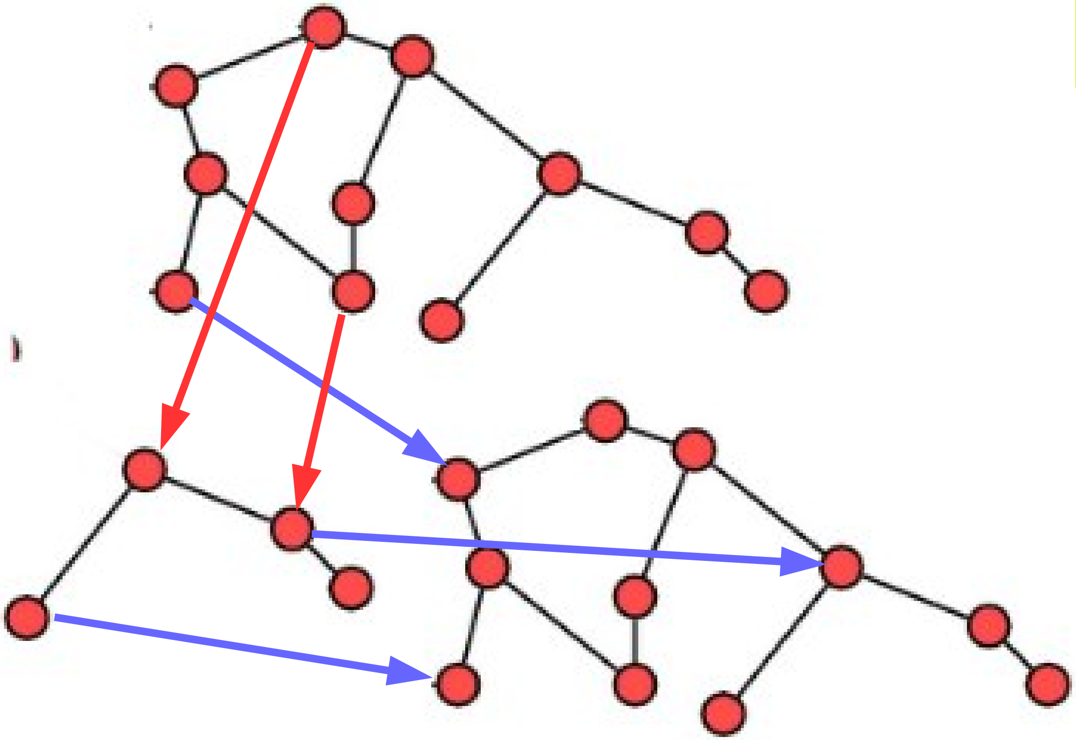


PHIS

- **Metadata / ontologies provide the meaning of data**
  → Link each data element to a controlled, shared, vocabulary and machine readable
  → Structure the data in a graph



PHIS

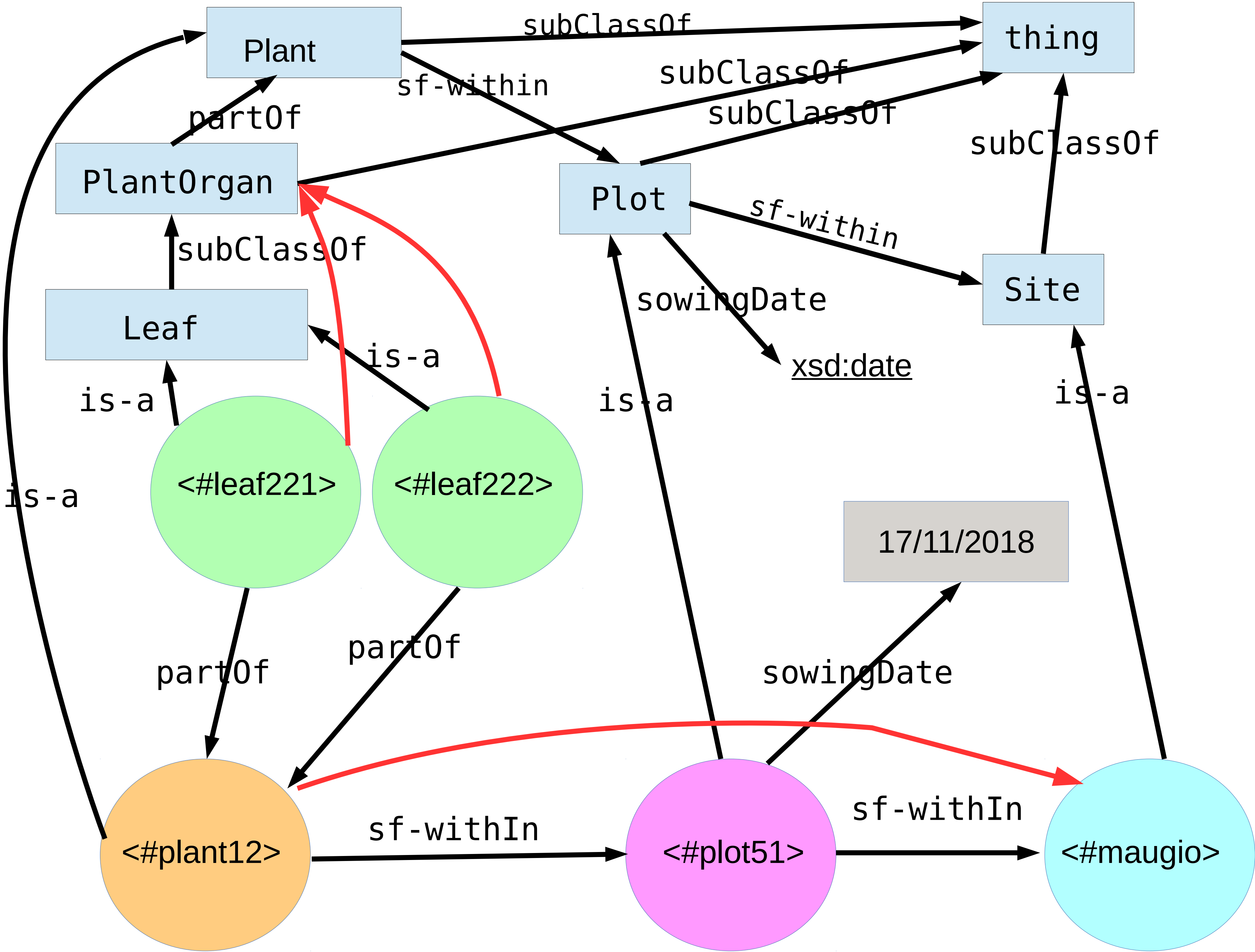


Reference ontologies
See AgroPortal

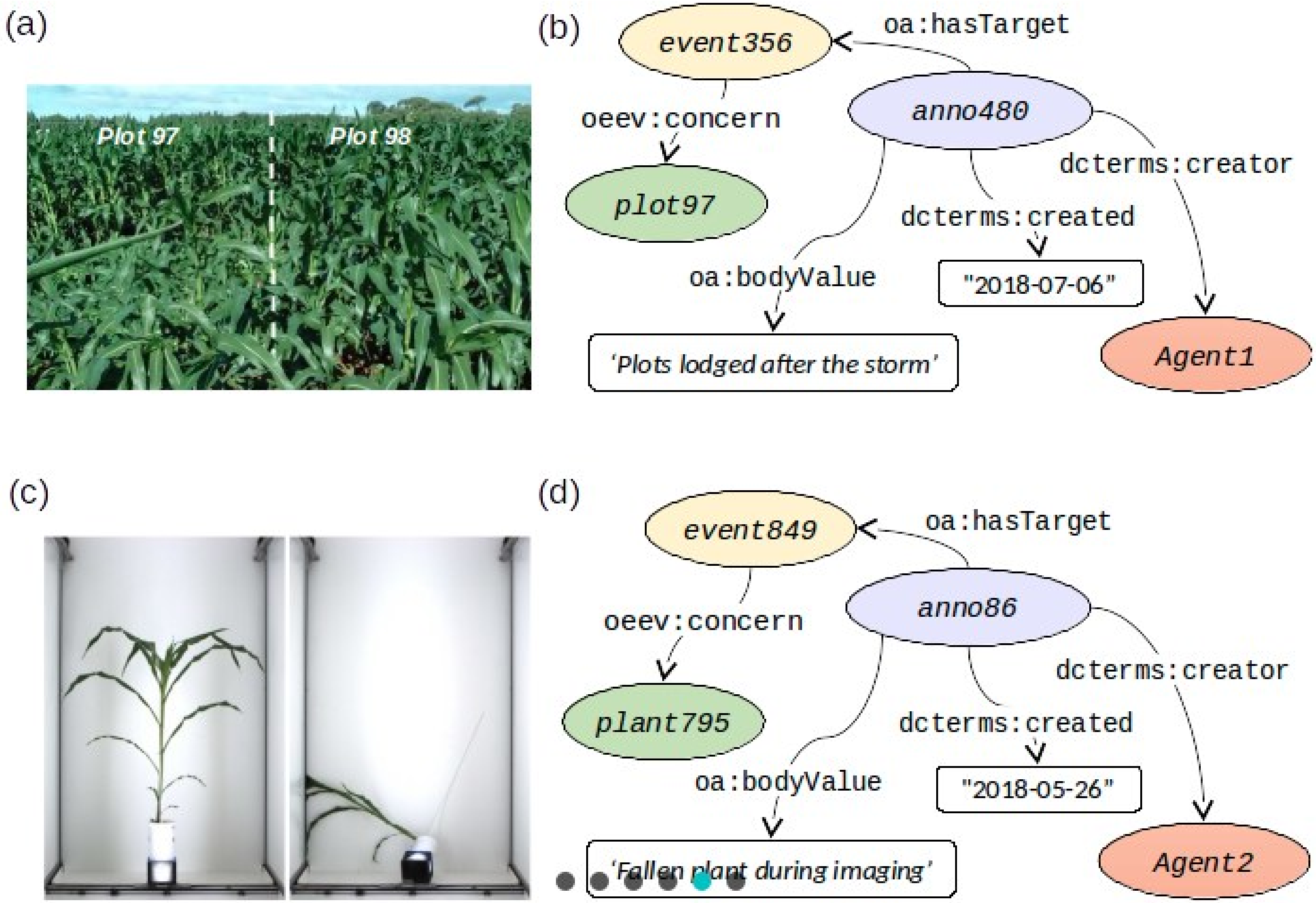
🟥

Application ontologies

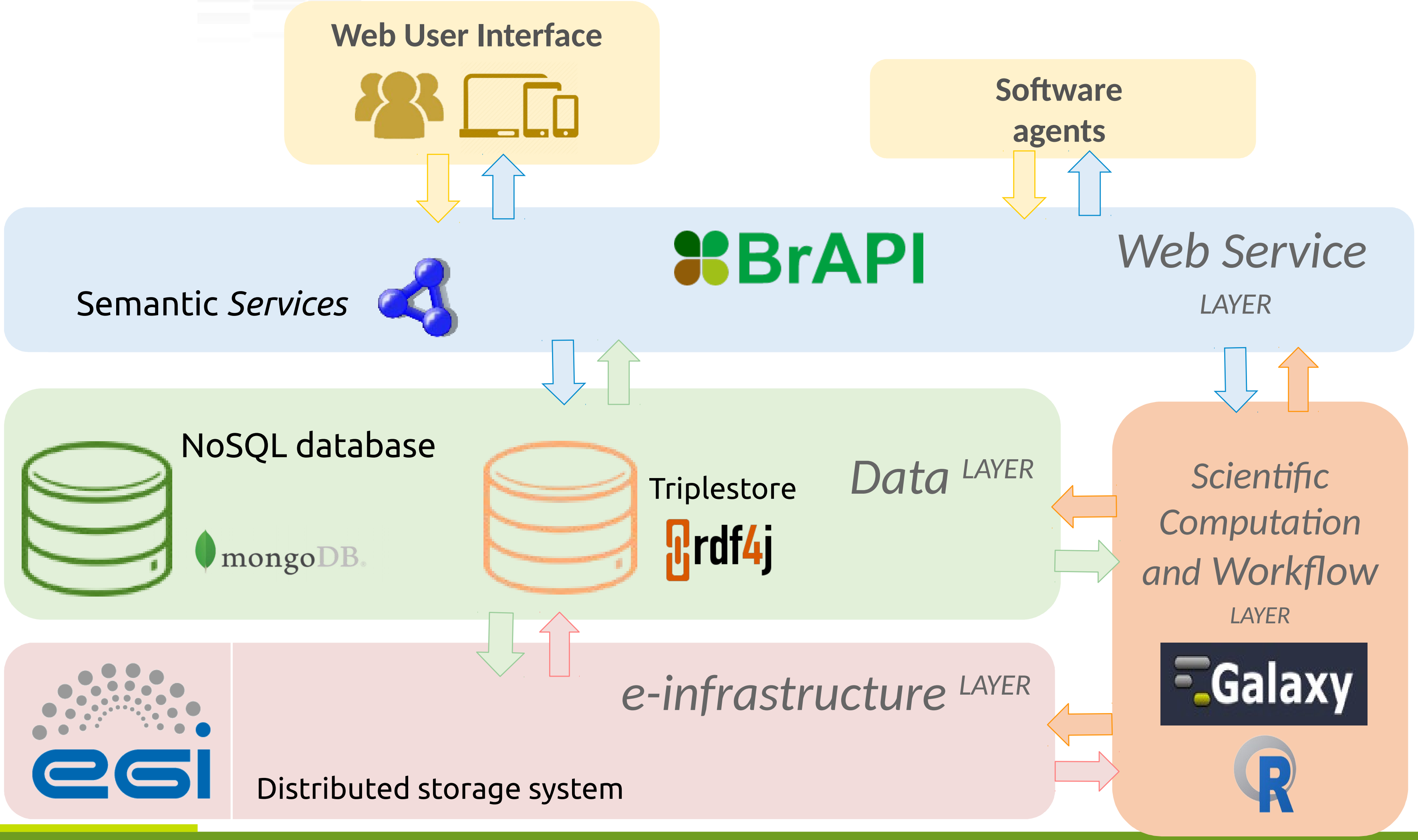
🟦


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

PHIS


(a)
Plot 97
Plot 98
(b)
event356
oa:hasTarget
anno480
oeev:concern
plot97
dcterms:creator
dcterms:created
"2018-07-06"
oa:bodyValue
'Plots lodged after the storm'
Agent1
(c)
(d)
event849
oa:hasTarget
anno86
oeev:concern
plant795
dcterms:creator
dcterms:created
"2018-05-26"
oa:bodyValue
'Fallen plant during imaging'
Agent2

OpenSILEX - PHIS



Web User Interface

Software agents

BrAPI

Semantic *Services*

Web Service LAYER

NoSQL database

mongoDB.

Triplestore

rdf4j

Data LAYER

Scientific Computation and Workflow LAYER

Galaxy

R

e-infrastructure LAYER

egi

Distributed storage system

Variable naming

What do we recommended in global context

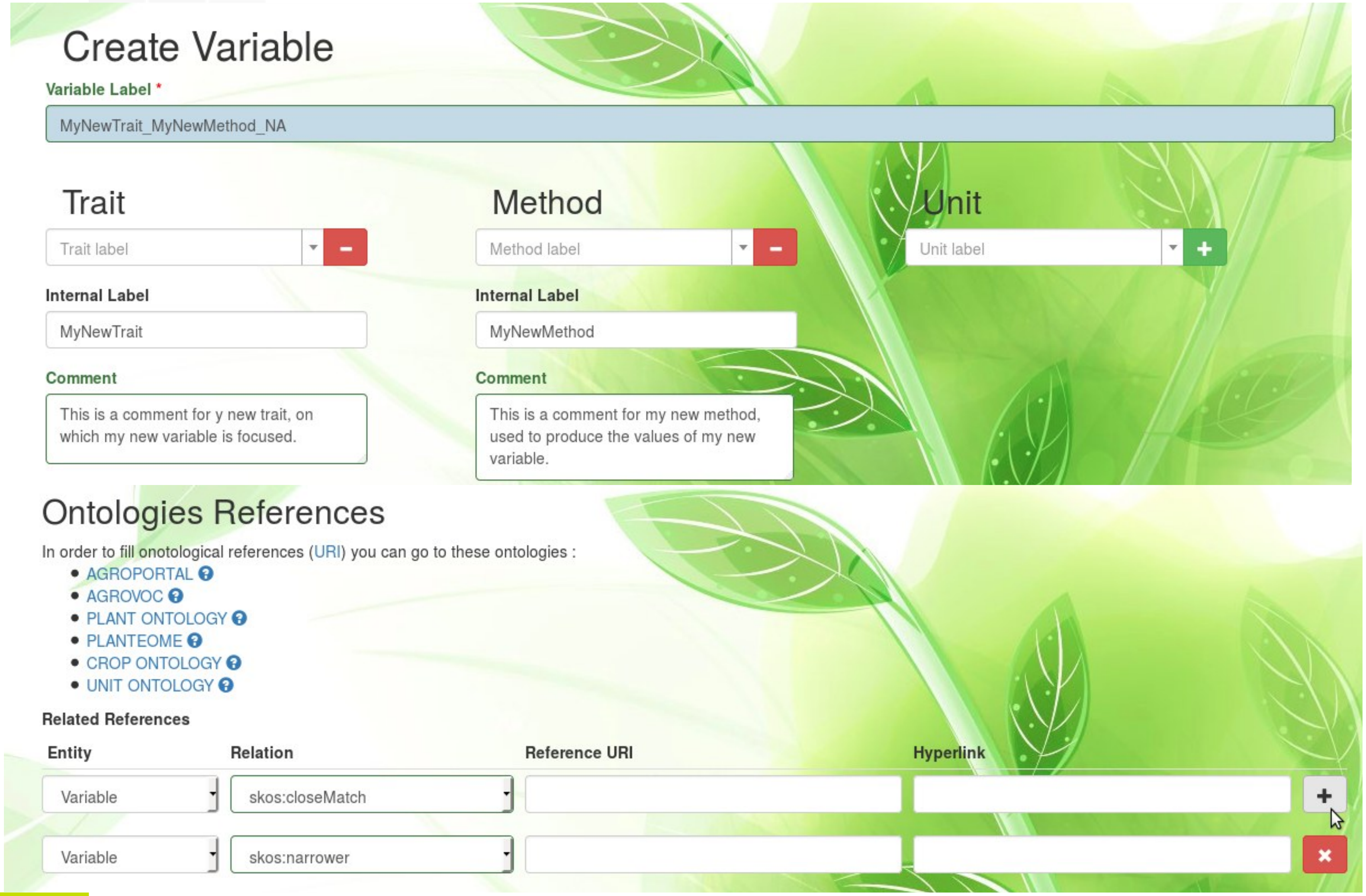
- Use URI for unambiguous name (in global context)
- Actionable URI for an accessible description of variable
  Description can be read by machine and human
- Try to reuse existing variable if available
- Use standardized/shared representation schema for formalisation of new variable (and share it)

  In PHIS

Variable = Entity + Quality/Quantity + Method + Unit

PlantHeight = plant + hauteur + ruler + cm

OpenSILEX - PHIS



Create Variable

Variable Label *

MyNewTrait_MyNewMethod_NA

Trait	Method	Unit
Trait label ▾ −	Method label ▾ −	Unit label ▾ +
Internal Label MyNewTrait	**Internal Label** MyNewMethod	
Comment This is a comment for y new trait, on which my new variable is focused.	**Comment** This is a comment for my new method, used to produce the values of my new variable.	

Ontologies References

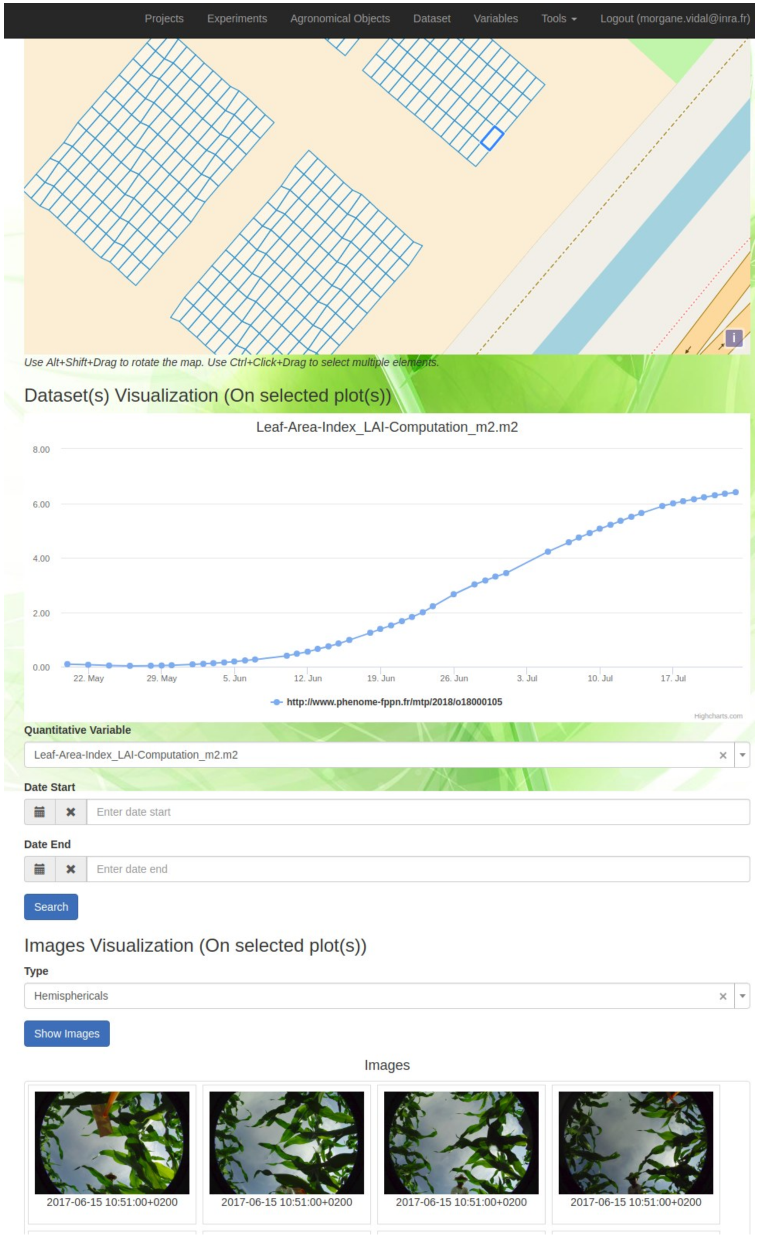
In order to fill onotological references (URI) you can go to these ontologies :

- AGROPORTAL ?
- AGROVOC ?
- PLANT ONTOLOGY ?
- PLANTEOME ?
- CROP ONTOLOGY ?
- UNIT ONTOLOGY ?

Related References

Entity	Relation	Reference URI	Hyperlink	
Variable	skos:closeMatch			+
Variable	skos:narrower			×

Trait – Provenance



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

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

Leaf-Area-Index_LAI-Computation_m2.m2

Quantitative Variable

Leaf-Area-Index_LAI-Computation_m2.m2

Date Start

Enter date start

Date End

Enter date end

Images Visualization (On selected plot(s))

Type

Hemisphericals

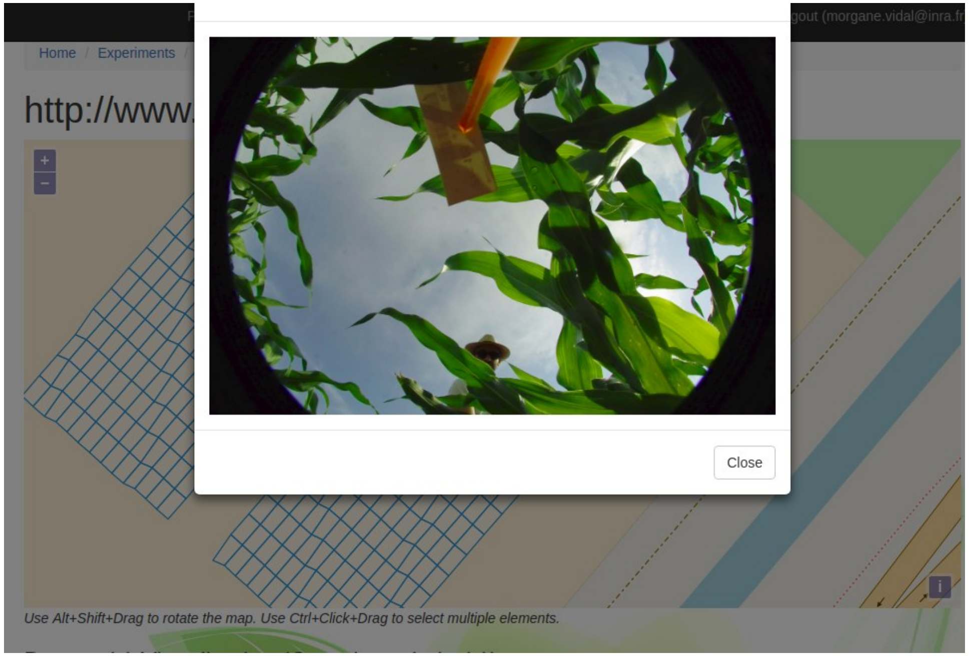
Images

2017-06-15 10:51:00+0200

2017-06-15 10:51:00+0200

2017-06-15 10:51:00+0200

2017-06-15 10:51:00+0200



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

PHIS

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

  → **distributed computing, distributed storage, backup**
- ✔**Flexible design**
- ✔**International identification (URI and DOI)**
- ✔**Semantic management (ontologies, standardized vocabularies)**
- ✔**Open technologies , Web APIs and portal interoperability**
- ✔**Provenance and reproducibility for data processing**
- ✔**Over 10 instances of PHIS** for various installations **(field and greenhouse)**
- ✔**Phenoarch instance → Over 700 Tb of data over 10 plant species**
- ✔**Other implementations of OpenSilex : WEIS, SunAGRI, SIUE**
- ✔**Open Software** - support and development (MISTEA team)

PHIS and OpenSilex

➢ **PHIS** demonstration

- http://phis.inra.fr/
- http://www.opensilex.org/opensilex-sandbox/web/
- Research paper: https://nph.onlinelibrary.wiley.com/doi/full/10.1111/nph.15385

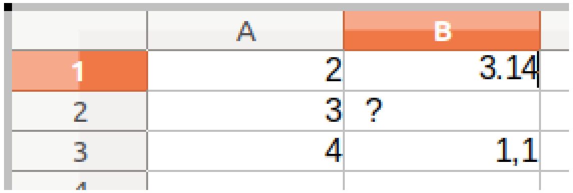
➢ How to contribute to OpenSILEX?

- 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/

I use spreadsheet! What I can do?



	A	B
1	2	3.14
2	3	?
3	4	1,1
4		

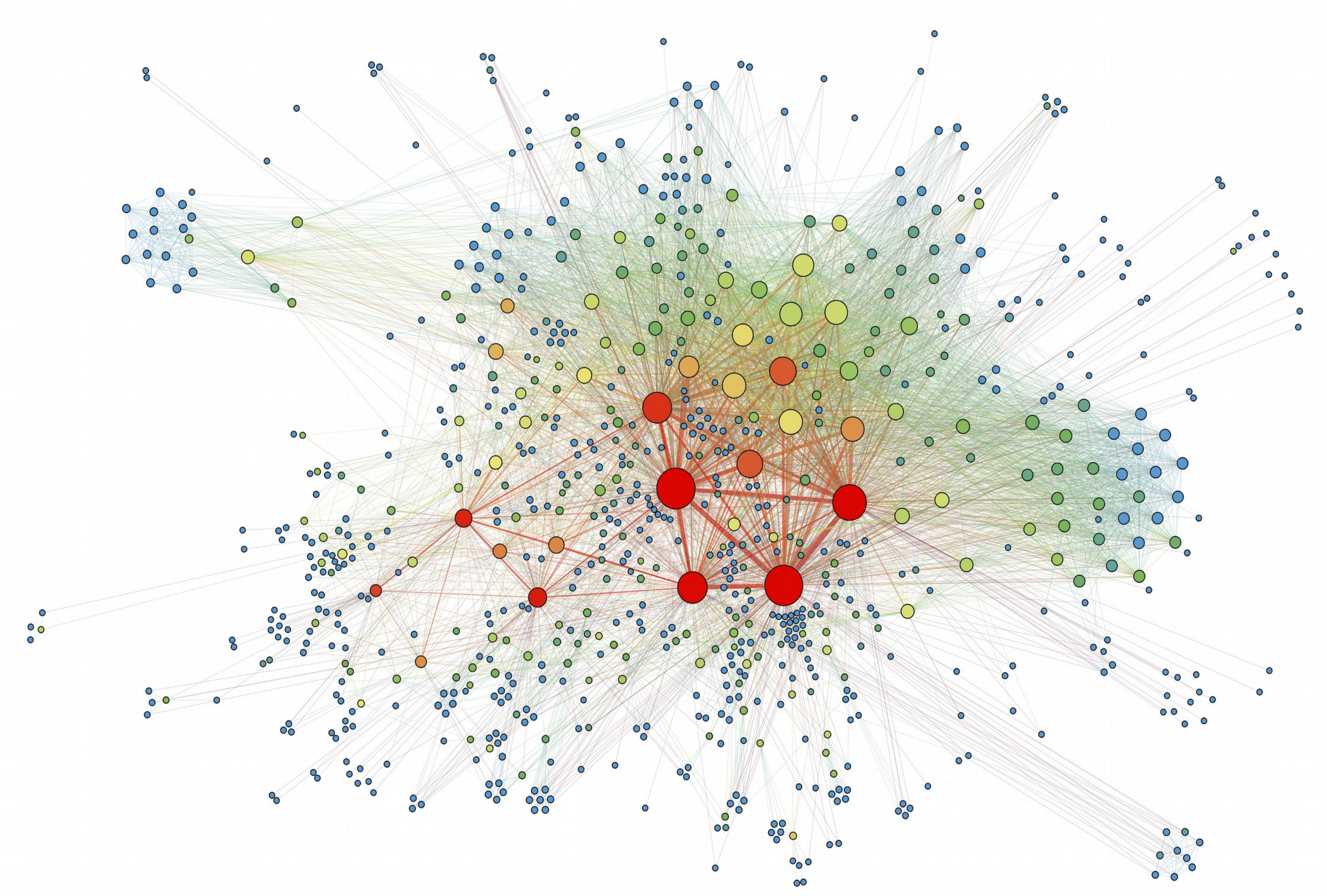
you can provide extra information known as metadata about CSV files using a JSON metadata file

2018ObservationSite.csv **then call the metadata file**
2018ObservationSite.csv-metadata.json

More information: http://w3c.github.io/csvw/primer/

- Merge duplicate resource IDs
- Reuse local unique
- Reuse local unique

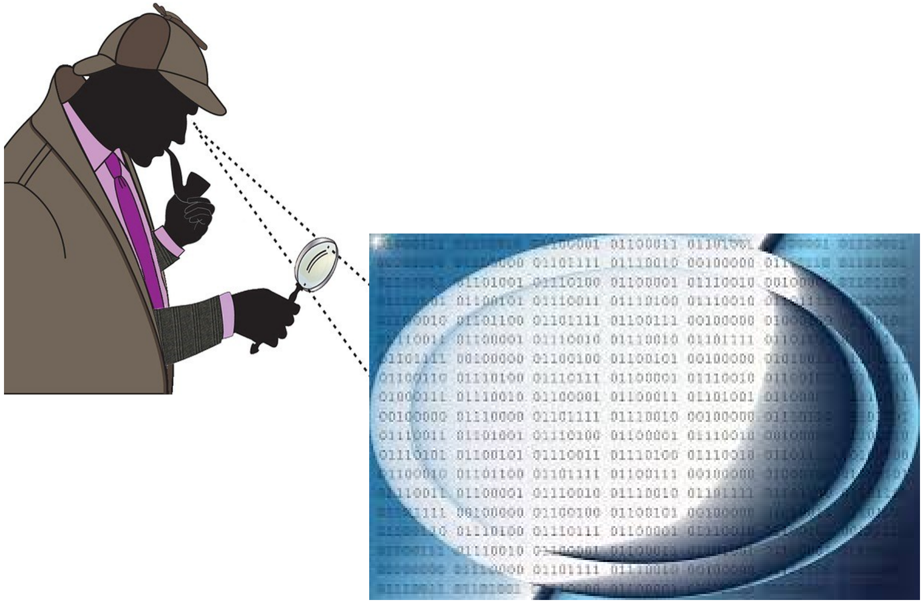
DATA in a Semantic GRAPH



Analysis of Graph Structured Data

How to structure data ?

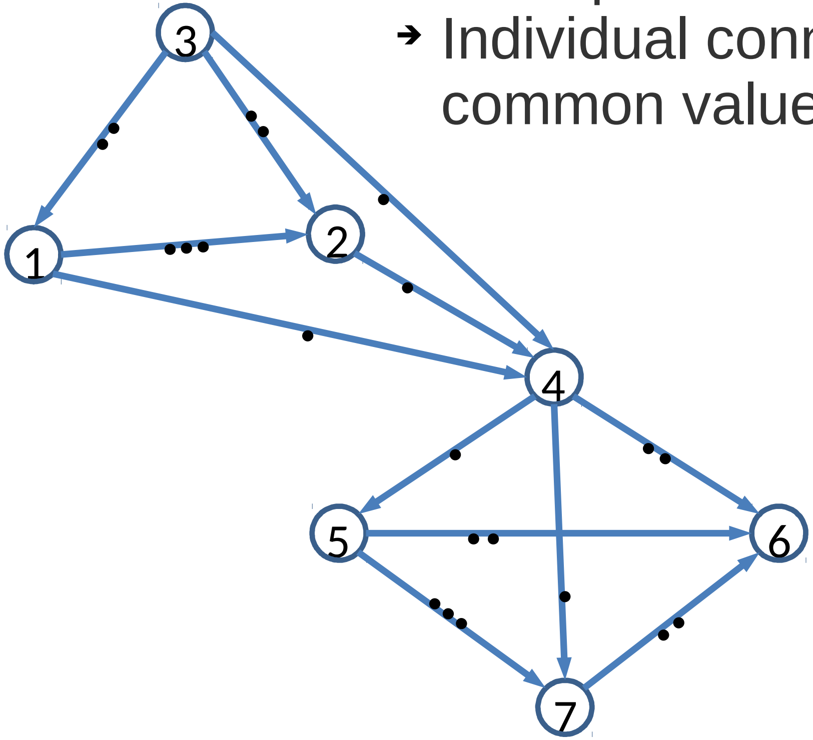
How to analyse data ?



Data analytics – Overview

Example: Set of variables

- Individual connections depend on the number of common values



Data analytics – Overview

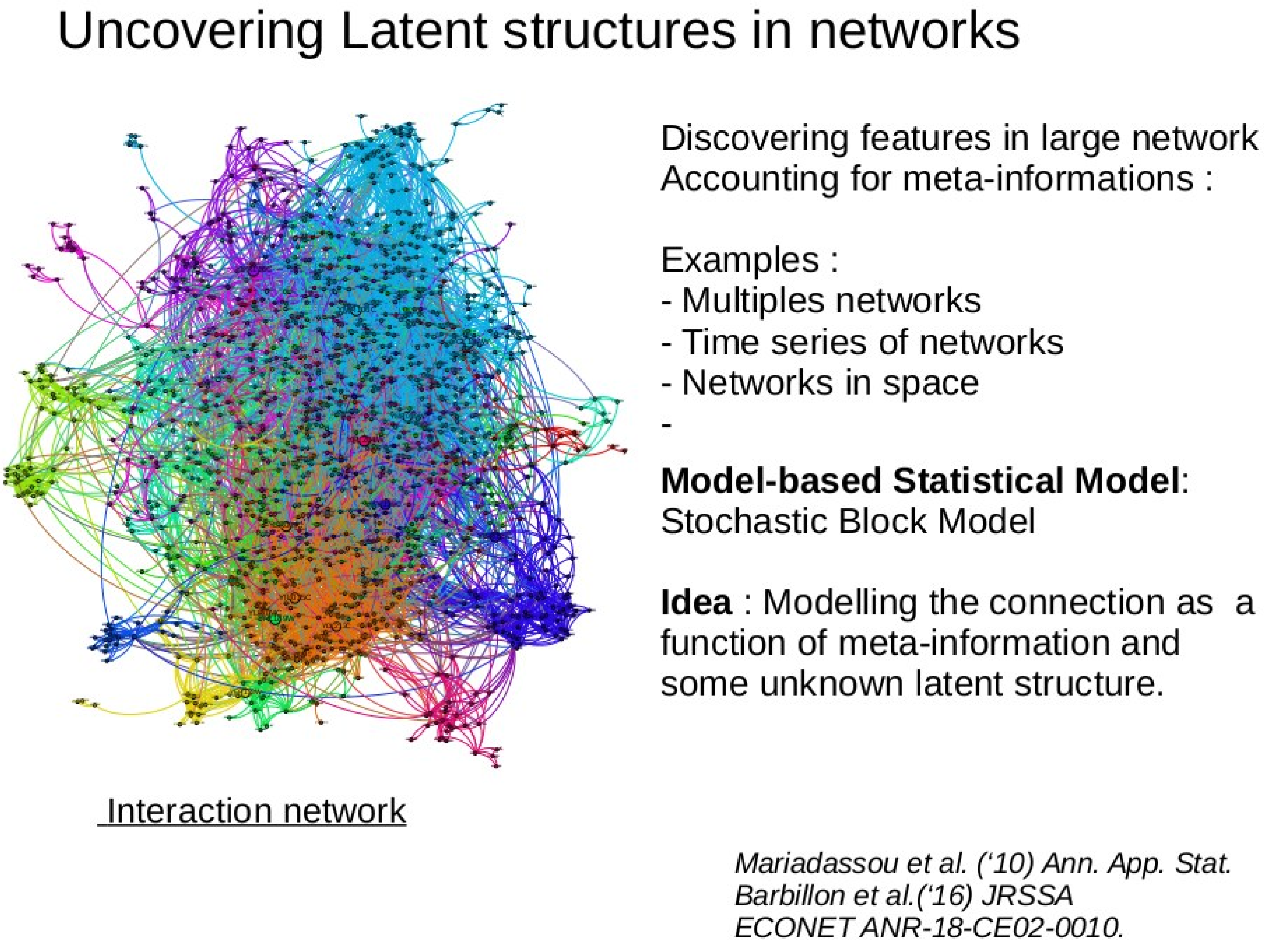
Objective: find combinations of variables (RDF properties) in heterogeneous large dataset that discriminate a resource.

- Key : a set of properties (variables) that uniquely identify individuals
- Idea: Use key Interpret numerical data in a symbolic way
- Automatically discover keys and evaluate their quality

- *How do we discriminate the wines??*

Quantiles	
	PH
Wine1	3.15
Wine2	3.22
Wine3	3.23
Wine4	3.24
Wine5	3.56
Wine6	3.68

Data analytics – Overview



Uncovering Latent structures in networks

Interaction network

Discovering features in large network
Accounting for meta-informations :

Examples :
- Multiples networks
- Time series of networks
- Networks in space
-

Model-based Statistical Model:
Stochastic Block Model

Idea : Modelling the connection as a function of meta-information and some unknown latent structure.

Mariadassou et al. ('10) Ann. App. Stat.
Barbillon et al.('16) JRSSA
ECONET ANR-18-CE02-0010.

Conclusion

- ✓ **Go to FAIR data Not Only for machine: allows teams and communities a better formalization of data**

- ✓ **Complex Big Data must be structured in graph in order to be able to FAIR, integrate and process it.**

- ✓ **Data mining and statistics methods are under development or already available to address volume and complexity issues**

- ✓ **Even if traditional methods remain accessible, the use of graph analysis can offer new ways**

Thank you for your attention