

Capitalisation et intégration sémantique de données de phénotypage

Contact: Pascal.Neveu@inra.fr





Data Challenges

More and more data!

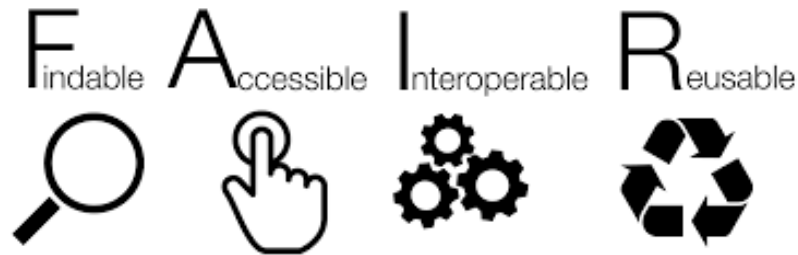
- Storage capacity, Network flow, etc.
1 Gigabyte : \$400K in 1980, \$10K in 1990, \$1K in 1995, \$10 in 2000, \$0,01 in 2017
- Various devices (on line or not), simulations, crowdsourcing, etc.
- Internet sources (Open, partners,)

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 !*
- ...

Need: A new generation of Information Systems



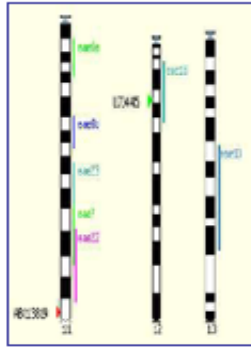
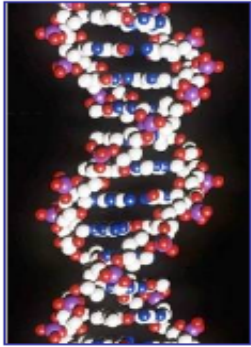
Findable: **PID**, standardized metadata and indexed in portals

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**

High Throughput Plant Phenotyping



High frequency observations
of trait dynamics
for big set of Phenotypes

Many Plant Genotypes

Interactions



Various Environments



High Throughput Plant Phenotyping: searching for the most adapted genotypes

Decision support

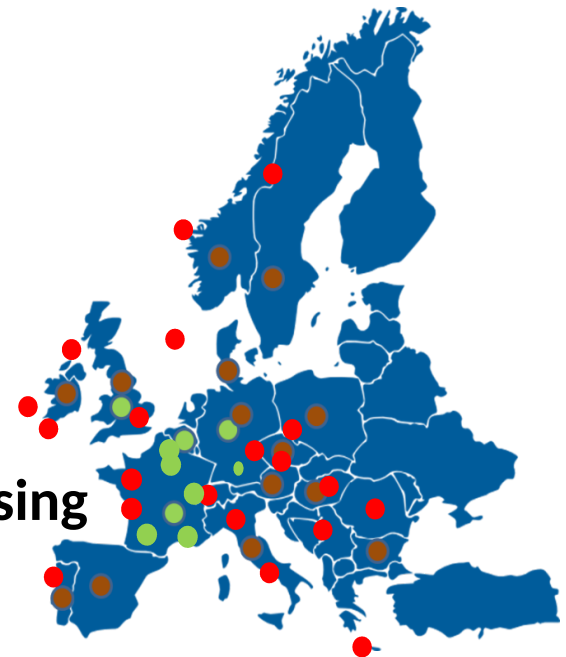
- Links genomics with plant ecophysiology and agronomy
- Phenotype-driven gene function discovery

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 European e-infrastructure

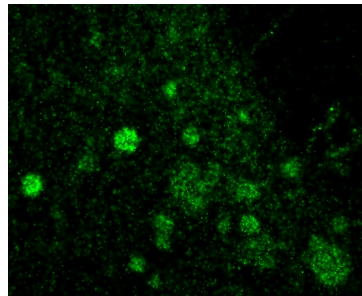
- Deals with several Petabytes of distributed data
- Makes FAIR data
- Based on Open technologies and standard (MIAPPE, BrAPI, etc)
- Standardized Identification
- Standardized Semantic
- Provenance and reproducibility data processing





Phenomics Data

Different scales



Intra-cellular



Organ



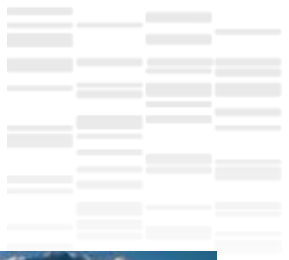
Plant



Field

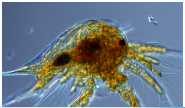


Region



Phenomics Data

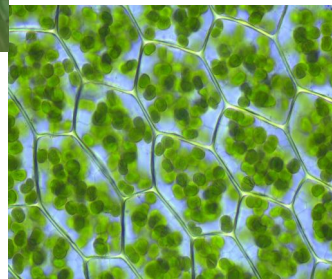
Different interactions





Phenomics Data

Different stages and transformations





Phenomics Data

From various contexts

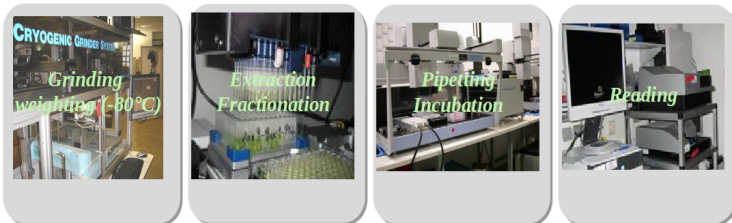
« omics » Platforms

Various data complex types

Genomics

Composition and the structure of biopolymers

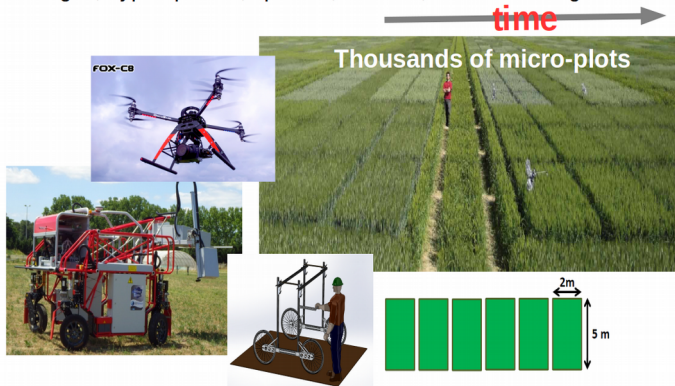
Quantification of metabolites and enzyme activities



Field Platforms

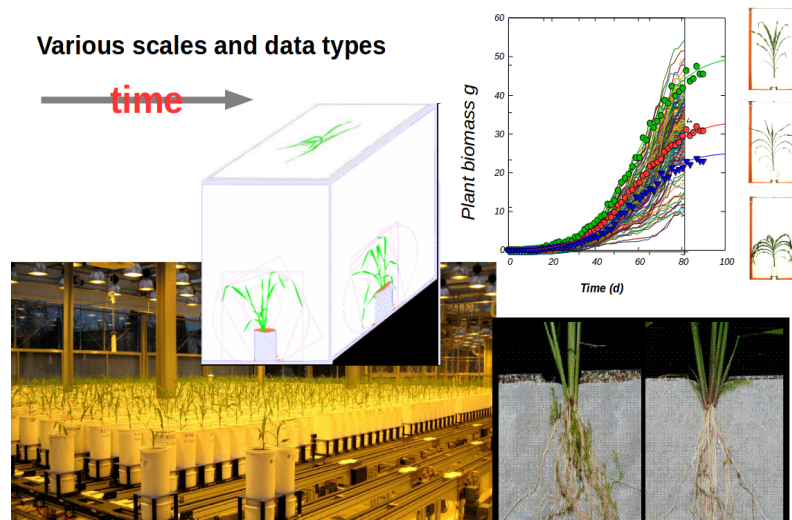
Various scales and data types

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



Green house Platforms

Various scales and data types



Farm Platforms

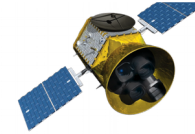
Various scales and data types from thousands of farms

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

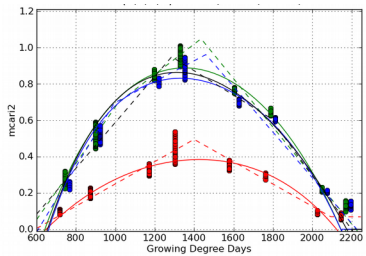
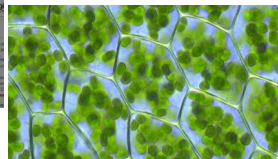
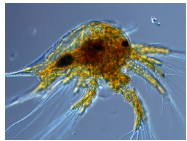
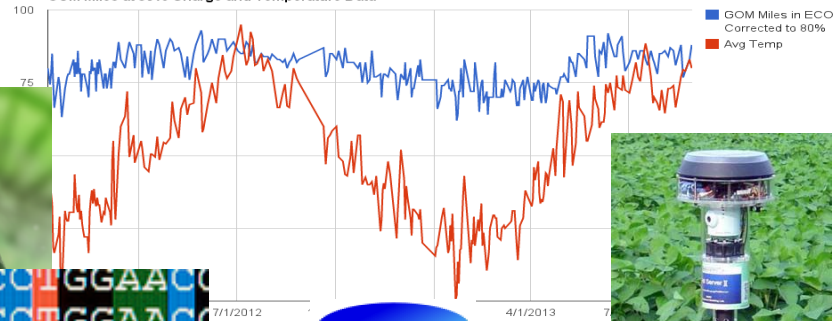




Phenomics Data



GOM Miles at 80% Charge and Temperature Data

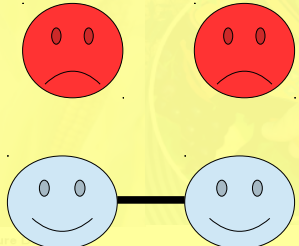
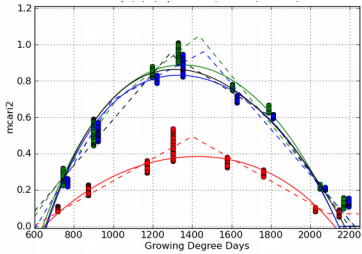
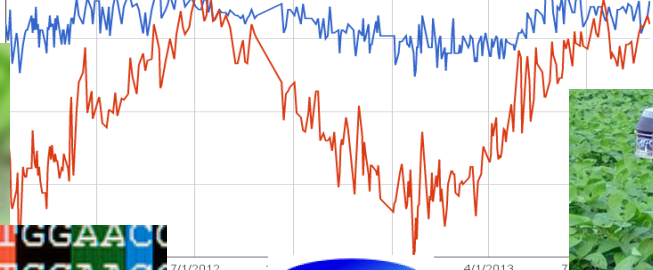
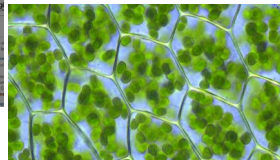
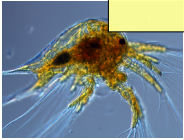




Phenomics Data



- Orphan data → Worthless!
- Data have value if they are grouped



Phenomics Data

How to structure data ?

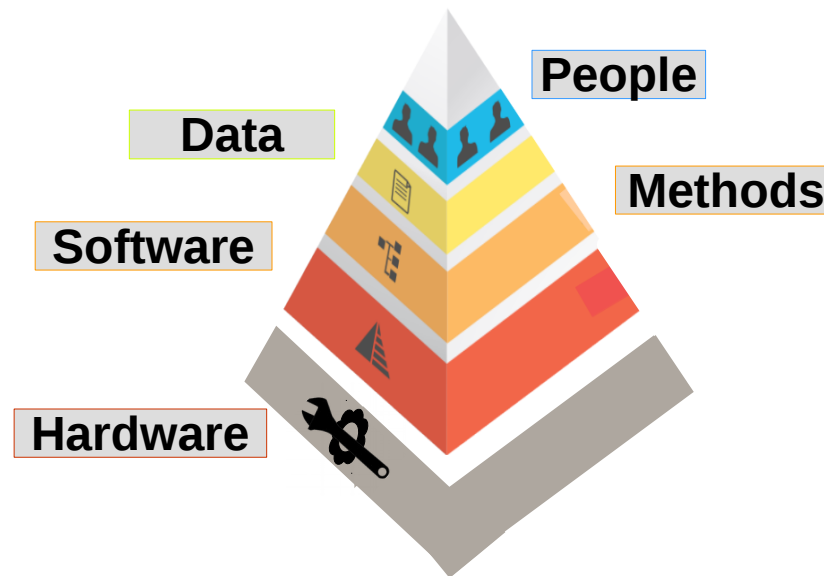




OpenSILEX

OpenSILEX is an Open source software set

- Methods, tools, components to implement information systems for experimental data in agriculture and environment
- for organisation, collection, structuration, storage, exchange and treatment of information





OpenSILEX - PHIS

- **PHIS is an instance of OpenSILEX**
- **Designed for data management in phenotyping platforms**
 - Management of huge, complex and heterogeneous data (millions of images, sensor data, from different sites, etc)
- **Implement good practices of data management**
 - Make FAIR data
 - Flexible
 - Ability to understand and reproduce data processing
 - Ability to enforce DMP and Open Science





OpenSilex approach

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

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

Events (management, faults, meteo, etc)

- Identified by **URI**

- **Organisation and linking** → objects and events with a controlled semantic (Ontology) such as a context specific application Ontologies (**RDF***, **OWL***) and allows to link reference ontologies (**SKOS***)

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

* **Semantic Web languages**

OpenSILEX-PHIS Identification

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

→ **Standardized, unambiguous**

`http://www.phis.inra.fr/path/identifiant`

Persistence and dereferencing (ePIC B2HANDLE)

Possible use of prefix

URI of plant :

`mp3:arch/2014/pl/000000012`

URI of pot :

`mp3:arch/2001/pt/000001542`

URI of cabin :

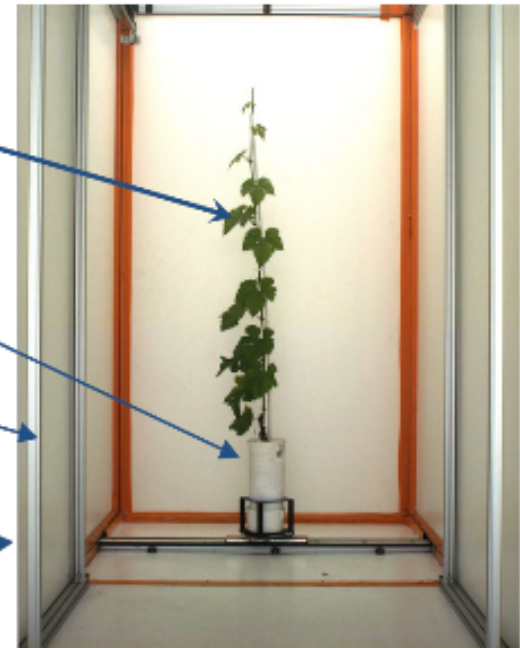
`mp3:arch/2010/ca/cabine2`

URI of camera :

`mp3:arch/2011/ss/00003312`

URI of image :

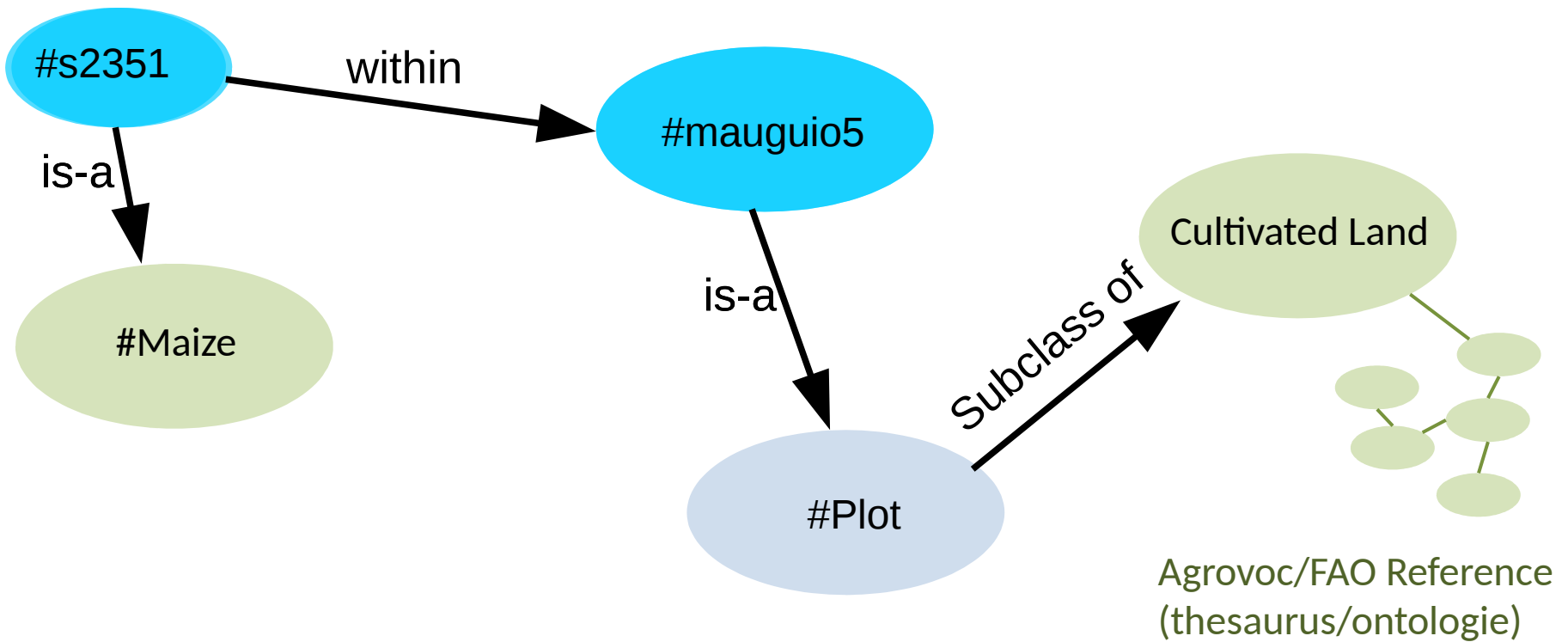
`mp3:arch/2015/im/000000564`





OpenSILEX - PHIS

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



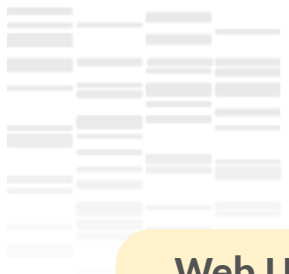
Agrovoc/FAO Reference
(thesaurus/ontologie)



OpenSILEX - PHIS

● Main technologies

- **Semantic Web** → semantic interoperability, complexity and metadata
- **NOSQL** for storage large data (spatial features)
- **Web Services** for data access and data publication
- **R interfaces** for data visualisation and data analytics



OpenSILEX - PHIS

Web User Interface

Software agents

Web Service LAYER

Semantic Services

Data LAYER

NoSQL database mongoDB.

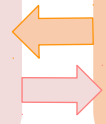
Triplestore **rd4j**

e-infrastructure LAYER

ESI Distributed storage system

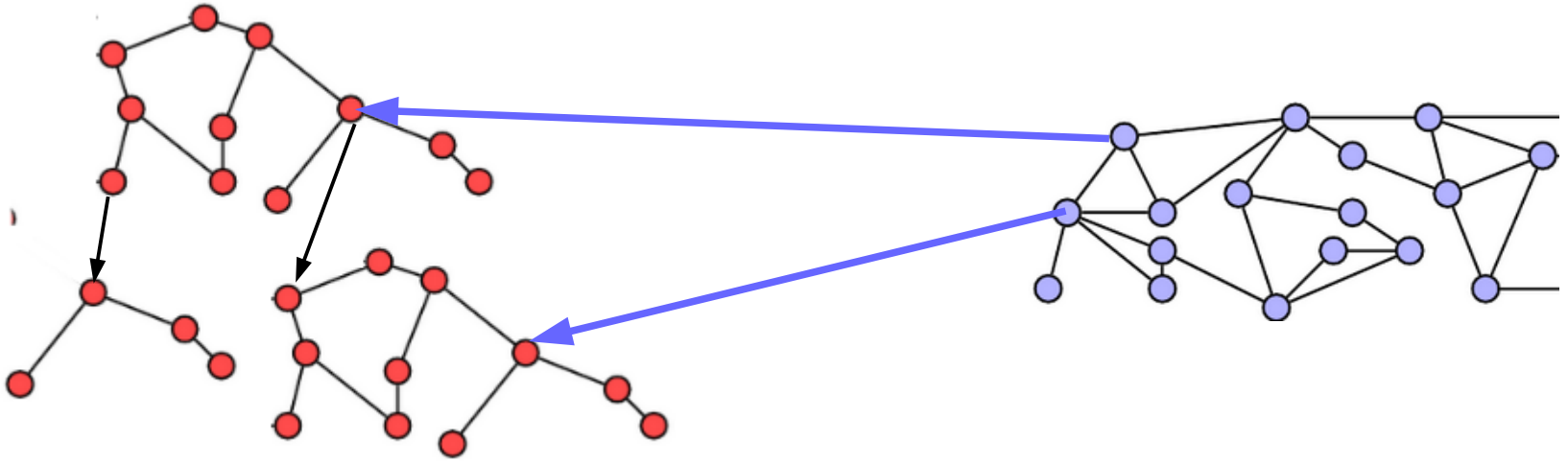
Scientific Computation and Workflow LAYER

Galaxy





OpenSILEX - PHIS



Reference ontologies

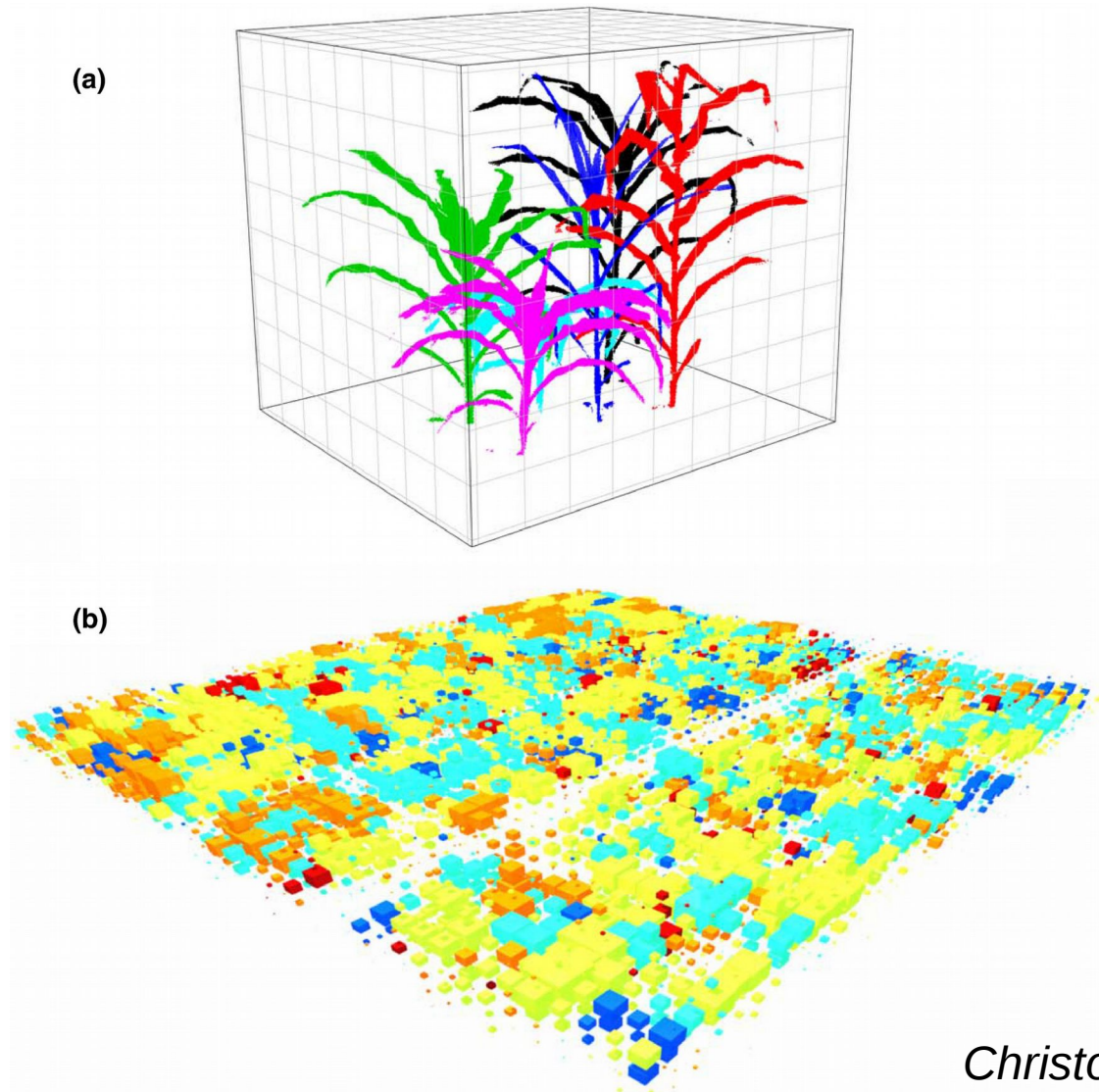
Application ontologies

Published in

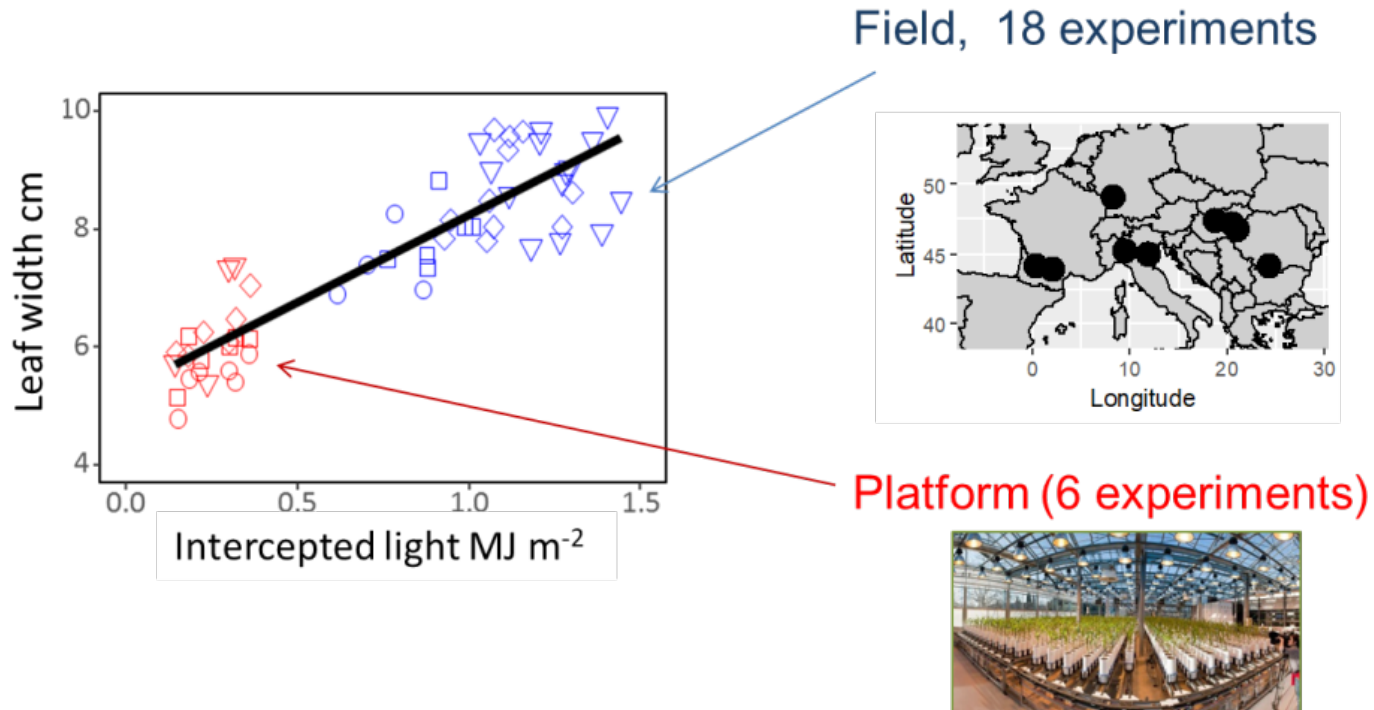


Knowledge Discovery Illustration

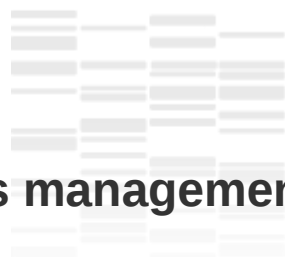
PHIS provides contextualisation: intercepted light value



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



OpenSILEX

- ✓ Allows management of huge and complex data
- ✓ Enables and facilitates cloud computing (data center, EGI)
 - distributed computing, distributed storage, backup
- ✓ Free software and Open technologies
- ✓ International identification (URI and DOI)
- ✓ Semantic management (ontologies, standardized vocabularies)
- ✓ Provenance and reproducibility for data processing
- ✓ Flexible design
- ✓ 5 instances of PHIS for various installations (field and greenhouse)
- ✓ Phenoarch instance → Over 300 Tb of data +10 plant species
- ✓ + 2 instances (WEIS, SunAGRI) + WUR, CIRAD, Univ of Tokyo
- ✓ MISTEA team: support and development, startup ?

PHIS

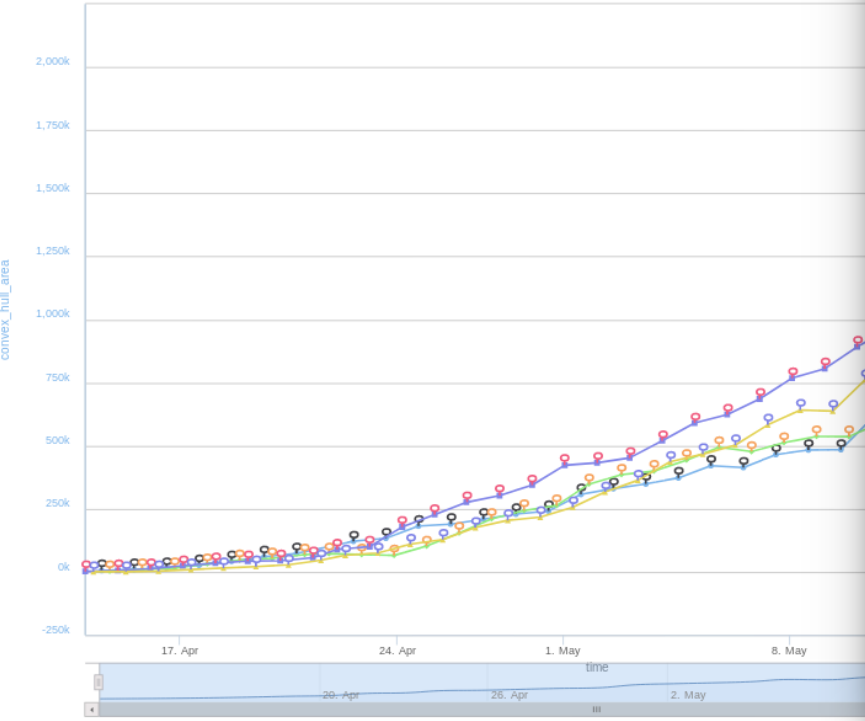
Event annotation

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/commi 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:

Projects Experiments Agronomical Objects Dataset Variables Tools Logout (morgane.vidal@inra.fr)



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

Search

Images Visualization (On selected plot(s))

Type
Hemisphericals

Show Images


Images



Trait – Images links

Home / Experiments

<http://www...>



Close

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



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

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

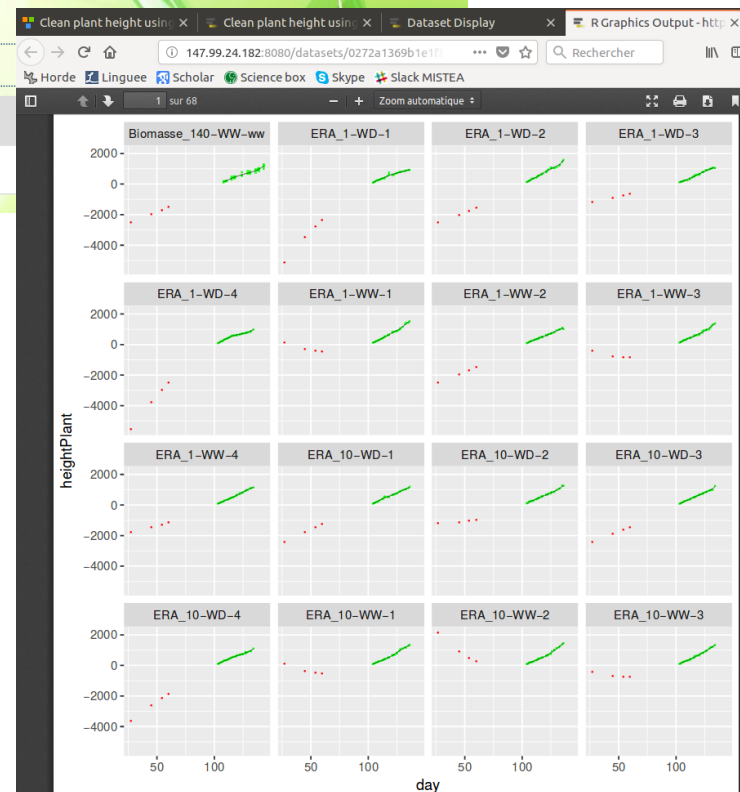
Invocation ID	40876639881ca029
History Id	6f91353f3eb0fa4a

The screenshot shows the Galaxy workflow interface. At the top, it displays the workflow name 'Clean plant height using default by Alamy' and the start time '13:53:00'. Below this, there are buttons for 'Switch to this history' and 'Show structure'. The main part of the interface is a list of workflow steps, each with a file size and a 'Show structure' button. The steps are:

- 11. Input: 2.8 MB
- 10. Log from plotting: 9.3 MB
- 8. Lines data: ~100,000 lines
- 7. Some data: 2.1 MB
- 6. log
- 5. Treated data: 4.1 MB
- 4. log
- 3. plotam file: 2.1 MB
- 2. 14 file
- 1. request file

The 'Input' step is expanded, showing a table with columns 'chr' and 'start'. The data rows are:

chr	start
1	139,28993480100
106	1000000000000
106	1000000000000
106	1000000000000
106	1000000000000





OpenSILEX

➤ **PHIS** demonstration

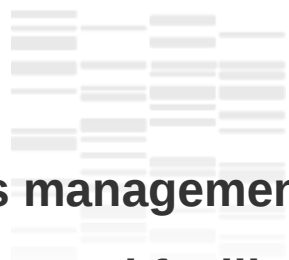
- <http://phis.inra.fr/> Or <http://www.opensilex.org/opensilex/web/>
Research paper:
<https://nph.onlinelibrary.wiley.com/doi/full/10.1111/nph.15385>

➤ 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/>



OpenSILEX

- ✓ **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 management (ontologies, standardized vocabularies)**
- ✓ **Portal interoperability and Open technologies**
- ✓ **Provenance and reproducibility for data processing**
- ✓ **Flexible design**
- ✓ **5 instances of PHIS for various installations (field and greenhouse)**
- ✓ **Phenoarch instance → Over 300 Tb of data over 10 plant species**
- ✓ **+ 2 instances (WEIS, SunAGRI)**
- ✓ **MISTEA team: support and development, startup ?**



Conclusion

OpenSILEX

- **Ensures and makes easy data findability and data access**
- **Provides description frame and an organisation**
- **Recommends and implements standards,**
- **makes easier data interoperability**
- **Provides data publication frame**