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 (online 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

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...

Farm Platforms
Various scales and data types from thousands of farms
- organ, plant, population, site
- Images, sensors, human readings...

Green house Platforms
Various scales and data types

Time

Thousands of micro-plots
Phenomics Data
Phenomics Data

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

How to structure data?
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/identifier

Persistency and dereferencing (ePIC B2HANDLE)

Possible use of prefix

<table>
<thead>
<tr>
<th>URI of plant:</th>
<th>URI of pot:</th>
<th>URI of cabin:</th>
<th>URI of camera:</th>
<th>URI of image:</th>
</tr>
</thead>
</table>
Metadata / ontologies provide the meaning of data
→ Link each data element to a controlled, shared, vocabulary and **machine readable** vocabulary
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

Semantic Services

NoSQL database
- mongoDB

Triplestore
- rdf4j

Data LAYER

Web Service

Software agents

Scientific Computation and Workflow LAYER

e-infrastructure LAYER

Distributed storage system

Web User Interface

Software agents

Web Service LAYER

Scientific Computation and Workflow LAYER

e-infrastructure LAYER

Distributed storage system
OpenSILEX - PHIS

Published in AgroPortal
PHIS provides contextualisation: intercepted light value
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 conditions.
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
Trait – Images links
PHIS
Variable management and interoperability

Create Variable

Variable Label
MyNewTrait_MyNewMethod NA

Trait
- Trait label
- Internal Label
MyNewTrait
- Comment
This is a comment for your 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:
- AGROPORTAL
- AGROVOC
- PLANT ONTOLOGY
- PLANTEOME
- CROP ONTOLOGY
- UNIT ONTOLOGY

Related References

<table>
<thead>
<tr>
<th>Entity</th>
<th>Relation</th>
<th>Reference URI</th>
<th>Hyperlink</th>
</tr>
</thead>
<tbody>
<tr>
<td>Variable</td>
<td>skos:closeMatch</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Variable</td>
<td>skos:narrower</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Trait</td>
<td>skos:exactMatch</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Method</td>
<td>skos:exactMatch</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
PHIS
Workflow management

Clean plant height using default

<table>
<thead>
<tr>
<th>_workflow name</th>
<th>Clean plant height using default</th>
</tr>
</thead>
<tbody>
<tr>
<td>Start</td>
<td>09-01-2018 13:53</td>
</tr>
<tr>
<td>End</td>
<td>09-01-2018 14:29</td>
</tr>
</tbody>
</table>

Open in Galaxy

Technical details

| Invocation ID        | 403876839881ea029               |
| History Id           | 669135316eb0fa4a                |

[Graphs and tables related to workflow management]
OpenSILEX

- **PHIS** demonstration

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

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