# Heterogenous plant data from data integration practices to analysis perspectives.

Intégration de données hétérogènes de plantes : des pratiques d'intégration aux perspectives d'analyse

Cyril Pommier / Heterogenous plant data from data integration practices to analysis perspectives.

Motivation: Increased sustainability of plant production is needed

- Less arable land
- Increasing world population

Source: FAO

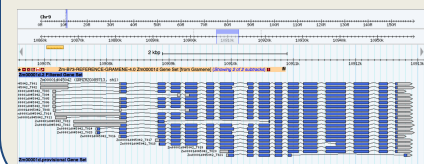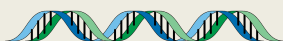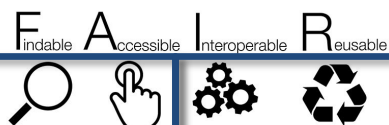- Climate change (extreme weather conditions, pests and diseases)
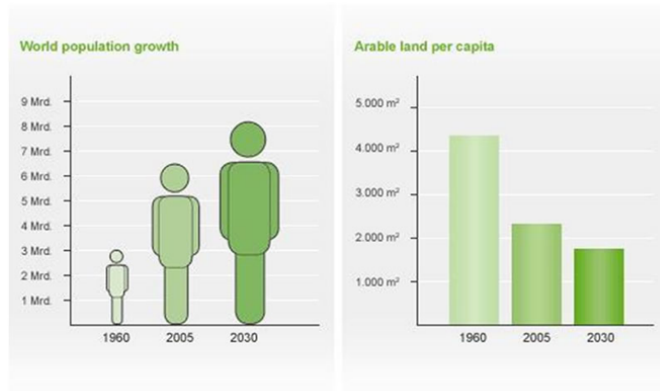
12 of the 17 UN Sustainable Development Goals (SDGs) address agriculture directly and indirectly

Genome / Environment / Phenome

Genetics & Genomics
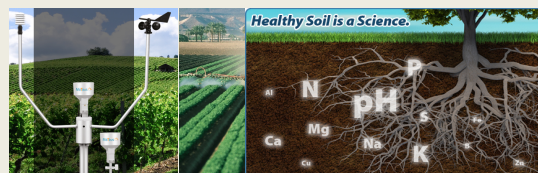
Homogenous (Mostly)
Central Access (Mostly)

F A I R
Findable Accessible Interoperable Reusable

Plant Breeding
Genetic variations by Traits

Climate Change Studies
Genotype by Environment
Phenology

Phenomics

Heterogenous
Distributed

Environment

Cyril Pommier / Heterogenous plant data from data integration practices to analysis perspectives.

Current Opinion in Plant Biology
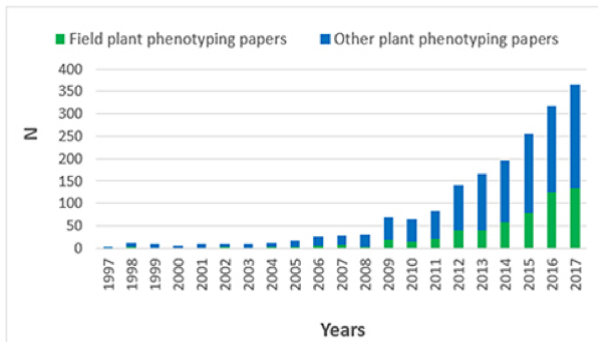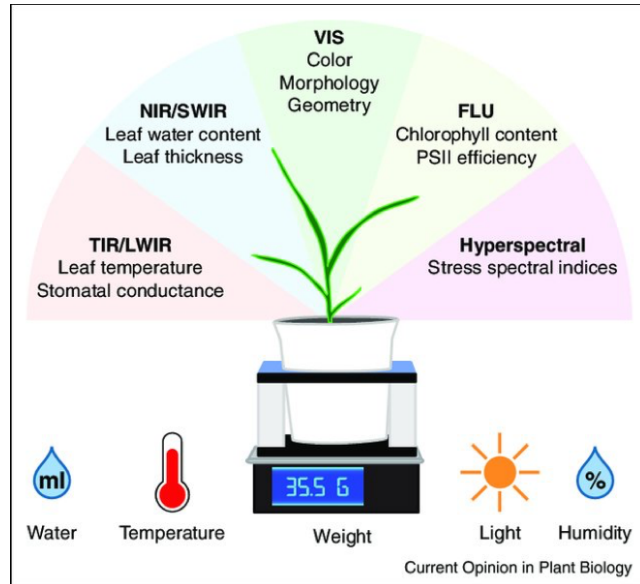
Heterogeneous and multidimensional

- Variability of experimental sites, designs, conditions
- Collected data - images, physical measurements, chemical assays, molecular biology assays, environmental,…

Dispersed

- No comprehensive public archive for phenotypic data

Poorly annotated

- Weaknesses in standards for data and metadata description

Plant 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*

Technical

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

Phenotype <u>Semantic</u> Standard: Ontologies

- Describing traits/features in specific plant species
- Crop Ontology Trait + Method + Scale Semantic model

Variable identification: Plant height example
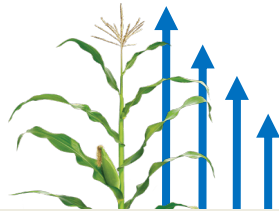
| Trait | **+** | Method | **+** | Unit |

M1: Total height

M2: First tassel branch

M3: Last expanded leaf

M4: Youngest growing leaf

...There is an uncountable number of combinations...
Each trait, method and unit has to be identified if we want to share and reuse data

T1: Plant Height

M5: Highest pixel corresponding to plant

U3: pixel

Slide from L. Cabrera-Bosquet

Phenotype <u>Structure</u> Standard
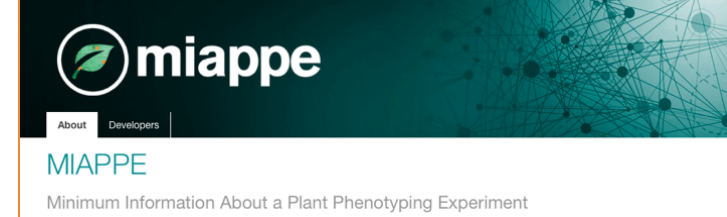
Minimal Information About Plant Phenotyping Experiment

www.miappe.org

- Many stakeholders
- Open Community
- Crops and woody plants
- Biologist friendly
  - **Clear definitions and examples**
  - **Excel templates**
  - **Trainings**
- Computer scientist friendly
  - **Explicit Data Model**
  - **ISA-Tools**
  - **Breeding API (BrAPI)**
  - **Phenotyping Experiment Ontology (PPEO) using OWL (http://agroportal.lirmm.fr/ontologies/PPEO)**
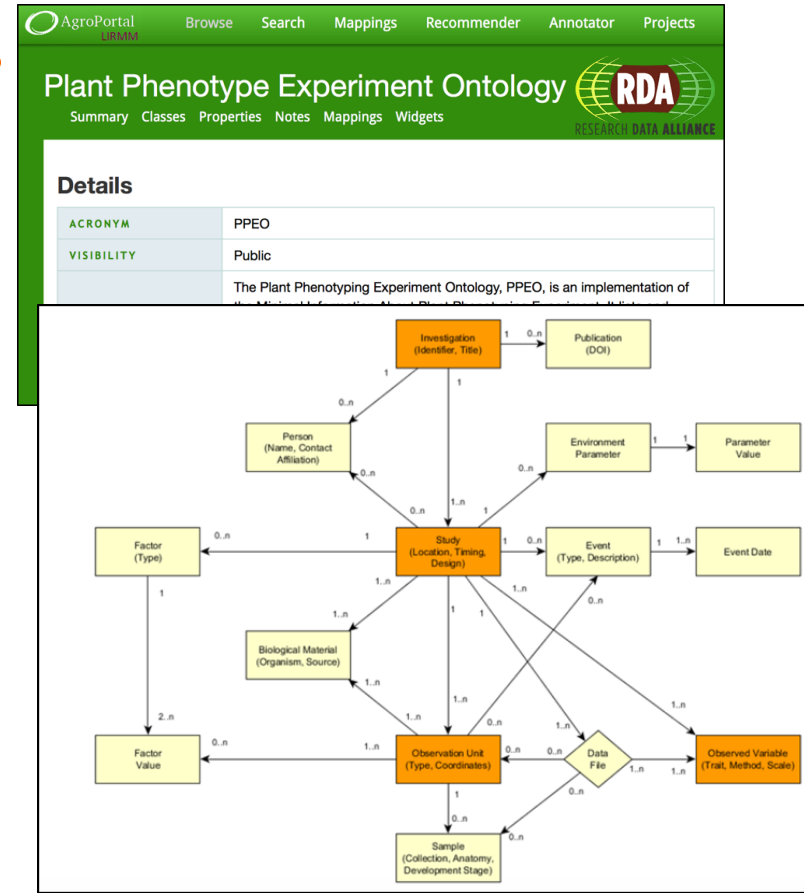
MIAPPE 1.1 Overview
Dataset

Example: Monclus *et al., 2012,* [2]

Populus deltoïdes
INRA:73028-62
doi:10.15454/1.4921786081024297E12

Data file

Data file

Data file

Data file
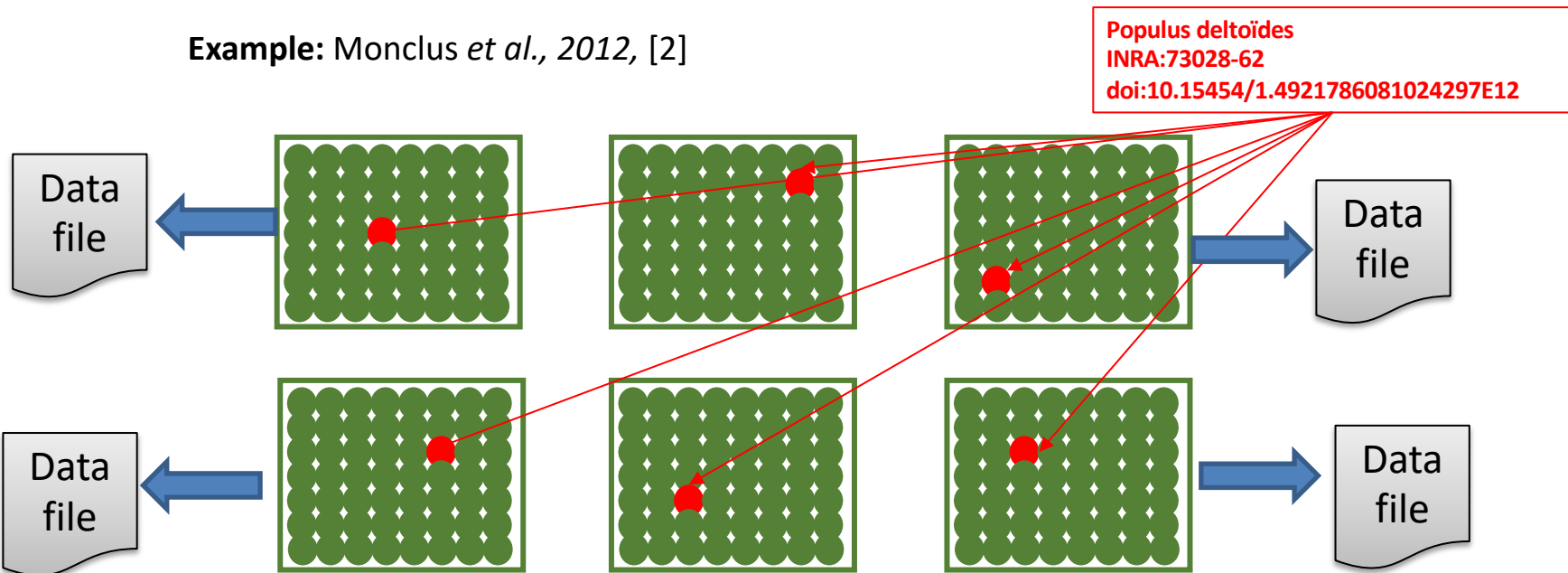
6 randomized blocks
1 observation unit = one tree
No treatment
6 replicates defined by their position in each block: row and column

MIAPPE V1.1 Overview

- **Investigation**: whole dataset
- **Study** : one experiment in one location for one to several year
- **Assay** *ie* **Observed Variable**: Level + Trait + Method + Scale/Unit
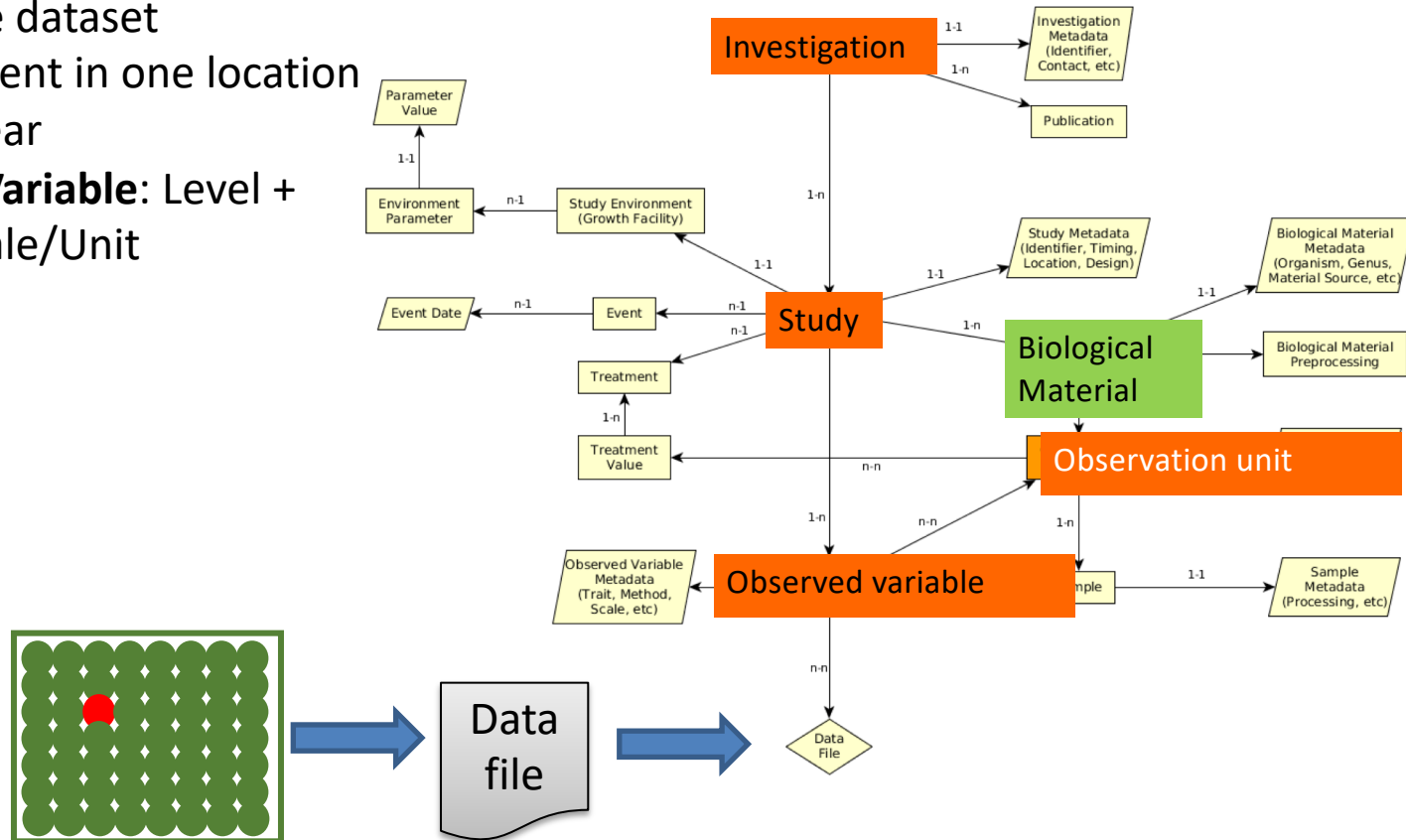
Plant Phenotyping Life cycle

Data acquisition →

- Traceability
- Raw measures
- Data Cleaning
- Platform IS (Emphasis IS, PHIS, …)
- Reproducibility
- Traceability
- Provenance

Data computation →

- New computation for each scientific question
- One raw dataset → many computed datasets

Data publication

- One Data Publication by datasets.
- **Platform IS**
  - Raw
- **FAIR Data Repositories**
  - Computed
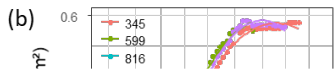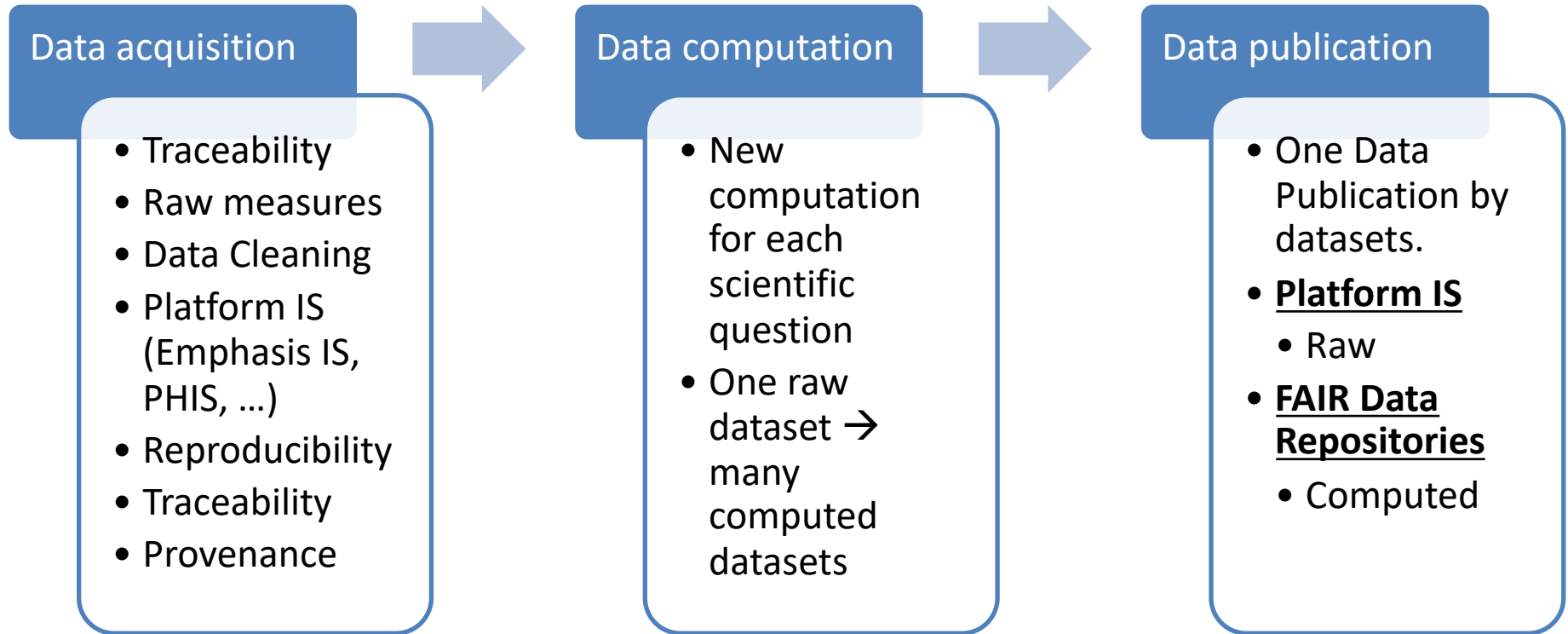
Genotype	traitement	Fusariose
pisson	low input	5
pisson	h	
Charger	low input	1
Charger	high N	2

Data → **Knowledge**

Variety charger

intensiv cultural practice

Cyril Pommier / Heterogenous plant data from data integration practices to analysis perspectives.

Data Integration
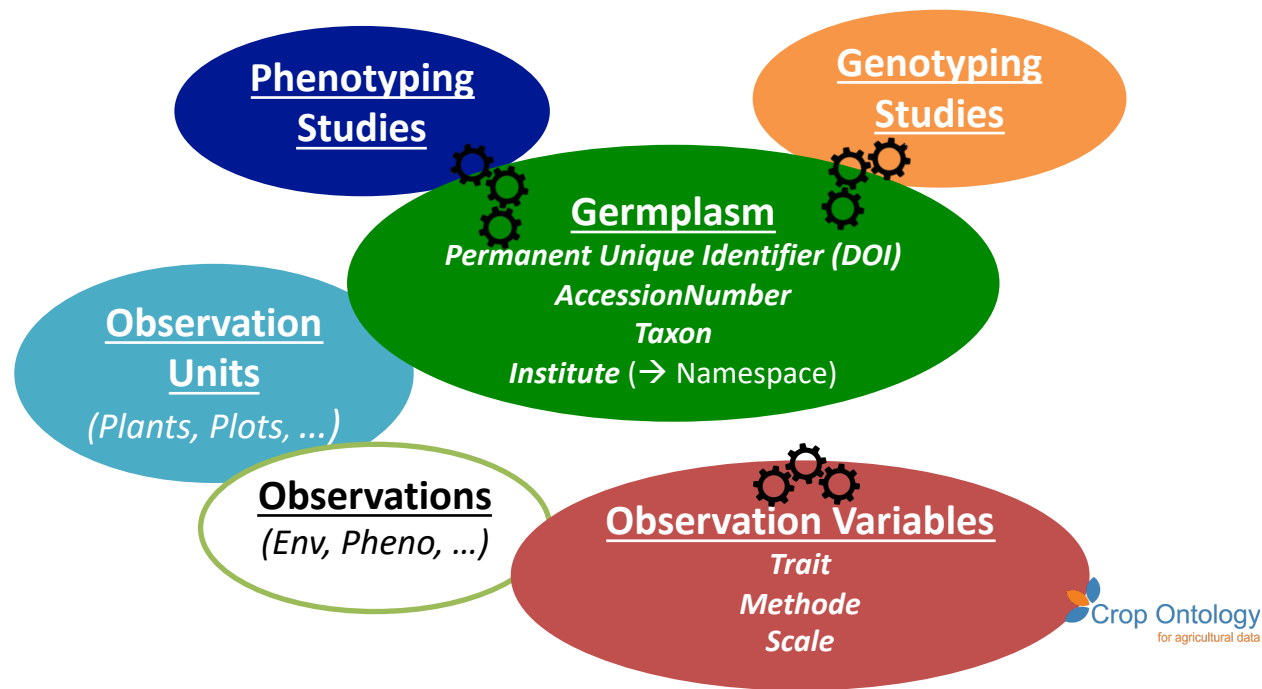
Phenotype <u>Technical</u> Standard

- Breeding API http://brapi.org/

- International collaboration

  - **Standard Open Web Service API, REST**

  - **Information Exchange, Main target: Breeding**

  - **Excellence in Breeding platform (CGIAR)**

- Major INRA & Elixir Contribution

  - **Phenotyping specification**

T3
CIRAD
GOBII
Wageningen
Cornell
iPlant

Selby *et al. Bioinformatics* (2019),
doi.org/10.1093/bioinformatics/btz190

Phenotype Technical Standard   ❖❖ BrAPI

- Connect data repositories and tools:
  - **Genotype visualization (Flapjack)**
  - **Studies graph preview and filtering**
    - **BrAPPS : Tools integrable in any BrAPI compliant System**
      - **https://www.brapi.org/brapps.php**
  - **R analysis environment**
  - **Field data capture**
  - **FAIR Data discovery → Elixir FAIDARE**
    - **https://urgi.versailles.inra.fr/faidare/**
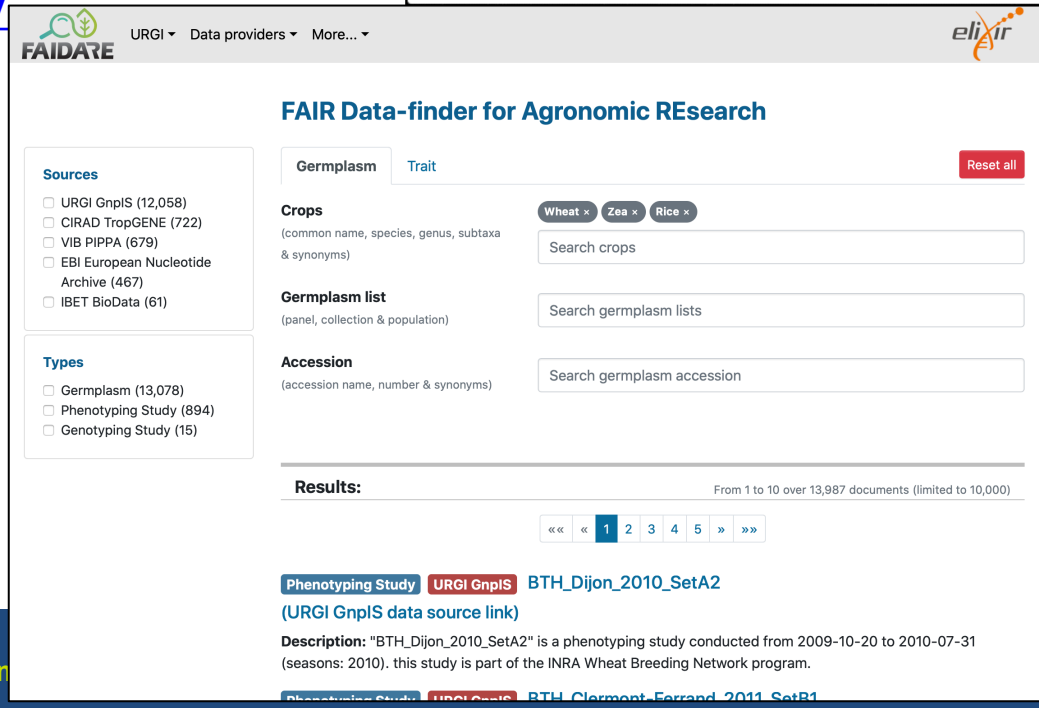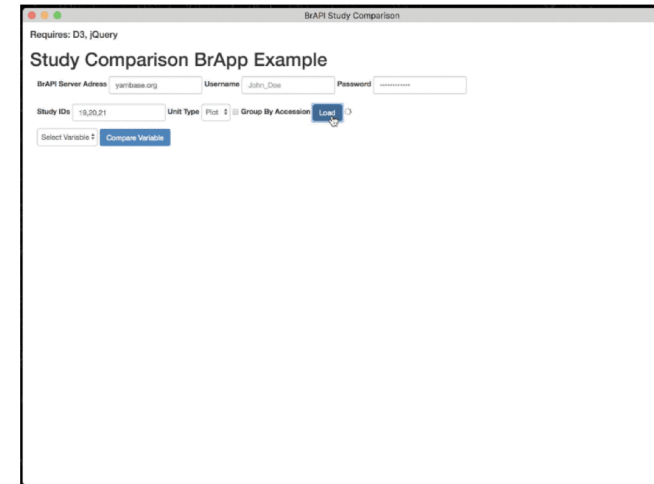
FAIDARE (FAIR Data-finder for Agronomic REsearch)

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

- Open plant data repositories federation
- How to join us
  - ◆ **Create your own BrAPI endpoint** (documentation: **https://brapi.org/**)
  - ◆ **Reference your BrAPI endpoint in FAIDARE** (documentation: **https://urgi.versailles.inra.fr/faidare/join**)
- Information & Support
  - ◆ **Elixir Plant Community : https://www.elixir-europe.org/communities/plant-sciences**
  - ◆ **INRA URGI: urgi-contact@inra.fr**

Create search criteria - Germplasm

Cyril Pommier / Heterogenous plant data from data integration practices to analysis perspectives.

Create search criteria - Trait

Find data

Cyril Pommier / Heterogenous plant data from data integration practices to analysis perspectives.

Filter data

Data Access

Cyril Pommier / Heterogenous plant data from data integration practices to analysis perspectives.

Data access

FAIDARE web services

https://urgi.versailles.inra.fr/faidare/swagger-ui.html

Cyril Pommier / Heterogenous plant data from data integration practices to analysis perspectives.

IA & New analysis tech for Data Integration

Variety charger is resistant to fusariose under intensiv cultural practice
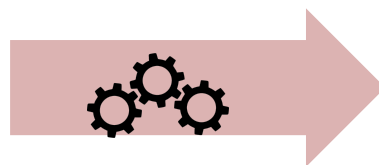
Charger
Triticum aestivum

Resistant under intesiv cutural practice

Fusariose
Fusariose
NA
NA

Phenotyping Studies

Genotyping Studies

Germplasm
Permanent Unique Identifier (DOI)
AccessionNumber
Taxon
Institute (→ Namespace)

Observation Units
(Plants, Plots,

Observations
(Env, Pheno, ...)

Observation Variables
Trait
Methode
Scale

M2 Master Internship on Text mining

- URGI / Bibliome partnership
- https://urgi.versailles.inra.fr/About-us/News/Text-mining-data-integration-web-devel-M2-internship

IA & New analysis tech for Data Integration

Charger
Triticum aestivum

Resistant under intesive cutural practice

Fusariose

Genotyping Studies

Phenotyping Studies

Germplasm
PUI (DOI)
AccessionNumber
Taxon

Observation Units
(Plants, Plots

Observations
(Env, Pheno, ...)

Obs Variables

Known analysis approaches
GWAS Analysis
Statistical Analysis
Images workflow
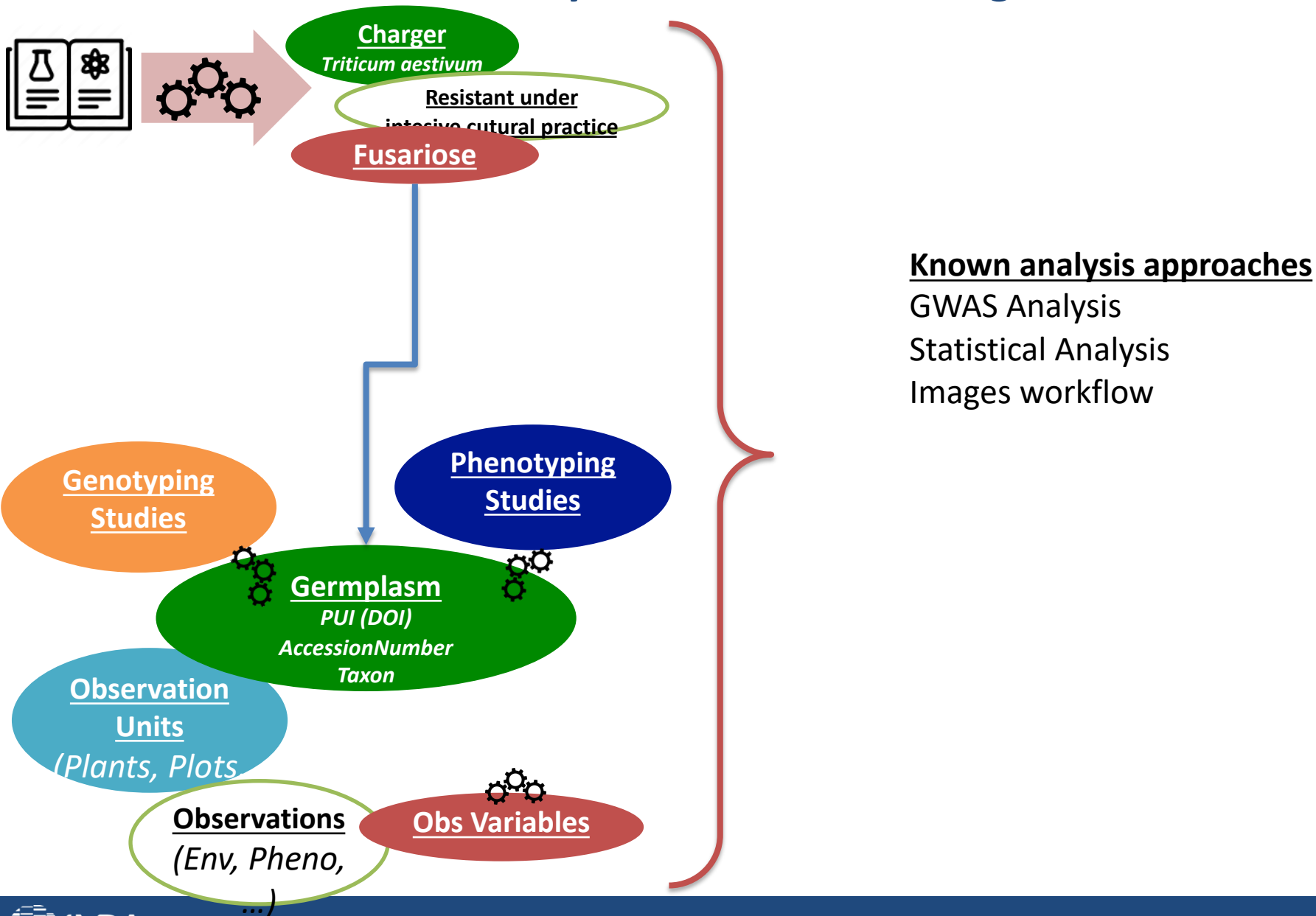
IA & New analysis tech for Data Integration
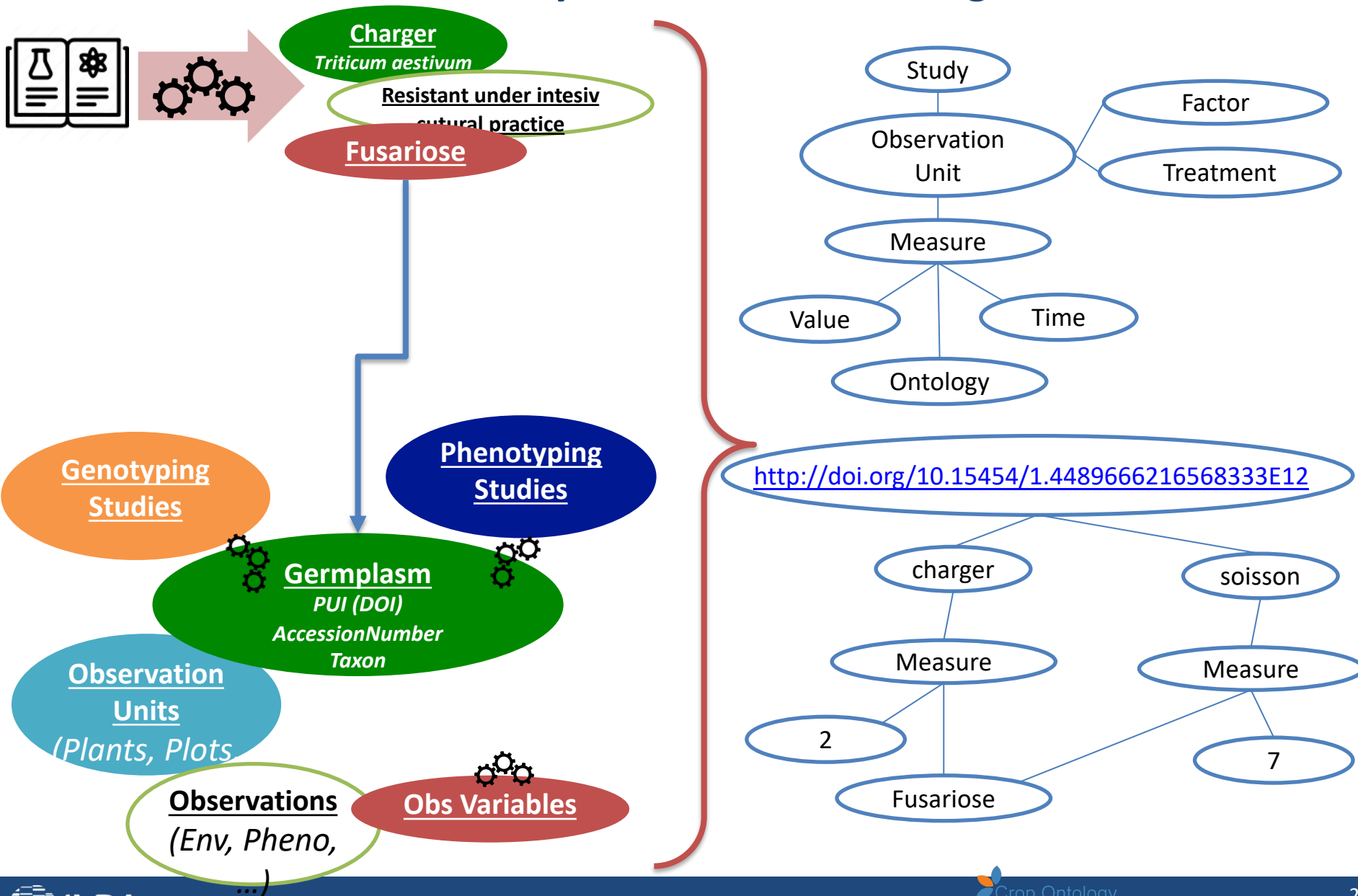
MIAPPE Semantic: RDA RDFENO

- MIAPPE OWL Implementation
  - **First draft:**
    https://github.com/MIAPPE/MIAPPE-ontology
- Brapi 2 MIAPPE RDF workflow
  - **http://ist.blogs.inra.fr/wdi/phenotypes-as-rdf/ (www.wheatIS.org data standards)**
  - **JSON LD based**
- Publication on MIAPPE.ORG
- Wheat dataset
  - **http://dx.doi.org/10.15454/1.4489666216568333E12**

Wheat Data Interoperability Guidelines

Home   Guidelines ⌄   Ontologies & Vocabularies   Use cases ⌄   Getting involved   About ⌄

Home / Phenotypes as RDF

Phenotypes as RDF

Phenotypes to Semantic web publication

The sharing and especially the reusability of complex data like phenotypes is a good use case for a publication following the semantic web principles. This page describes how to generate a Phenotype RDF dataset from a Breeding API MIAPPE compliant endpoint. The growing adoption of this API will facilitate the reusability of such a transformation process.

Recommendations

The RDF publication relies among other things on the Permanent Unique Identifier (PUI), either URIs or DOIs

Virtuoso SPARQL Query Editor

Default Data Set Name (Graph IRI)

Query Text

```
PREFIX brapi:  <https://brapi.org/rdf/>
SELECT ?network ?study ?attribute ?value
FROM <urn:urgi:pheno-brapi-inra-small-grain-cereals-network>
WHERE {?study a brapi:Study;
               ?attribute ?value.
        ?trial a brapi:Trial;
               brapi:hasName ?network.}
ORDER BY ?network ?study
```

(Security restrictions of this server do not allow you to retrieve remote RDF data, see details.

Results Format: HTML
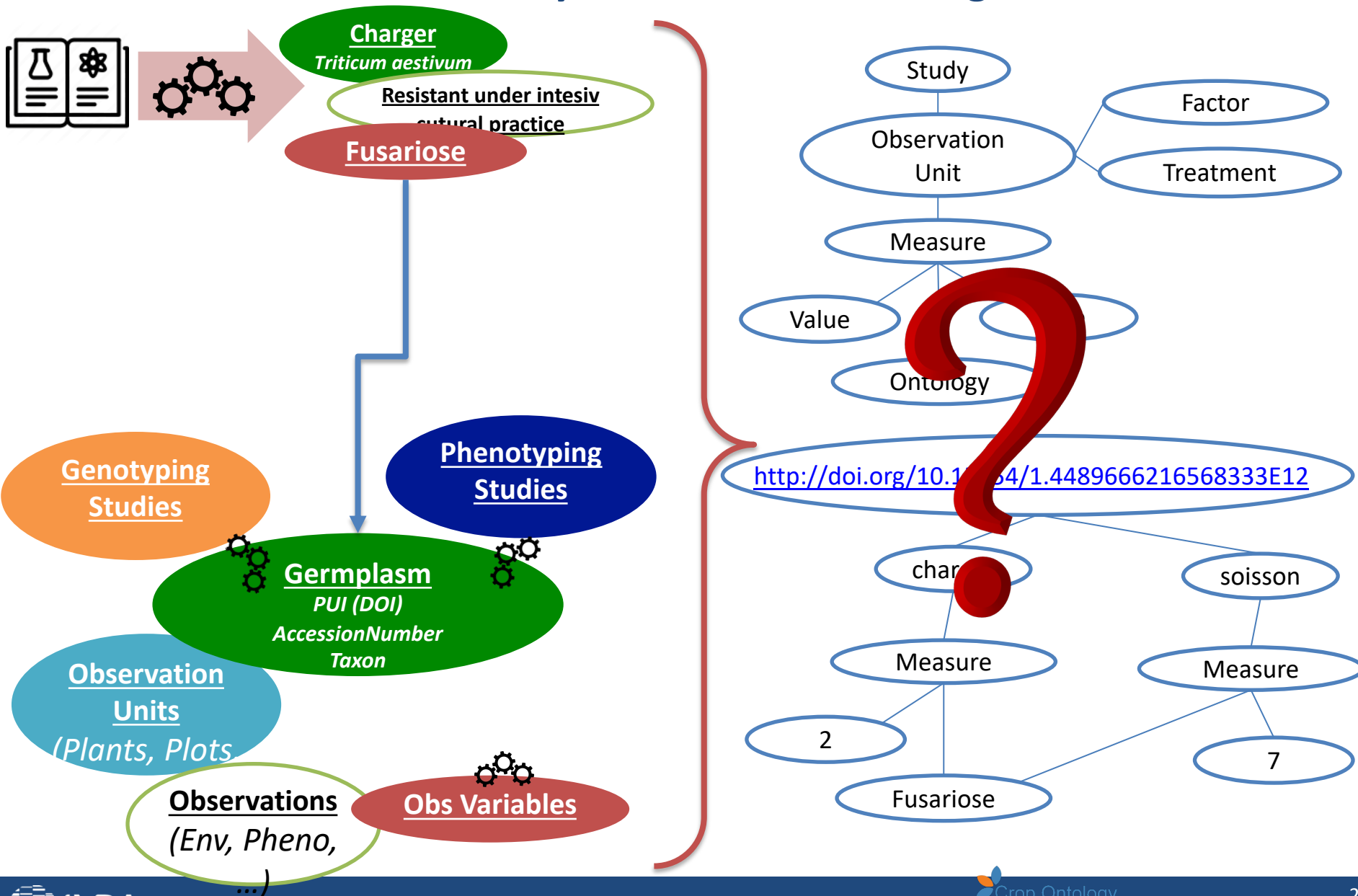Execution timeout: 0   milliseconds (values less than 1
Options: ☑ Strict checking of void variables   ☐ Log debug int

(The result can only be sent back to browser, not saved on the server, see details)

Run Query   Reset

IA & New analysis tech for Data Integration

THANK YOU