EMPHASIS

European Plant Phenotyping Network EPPN 2020

PHENOME French Plant Phenomic Infrastructure EMPHASIS

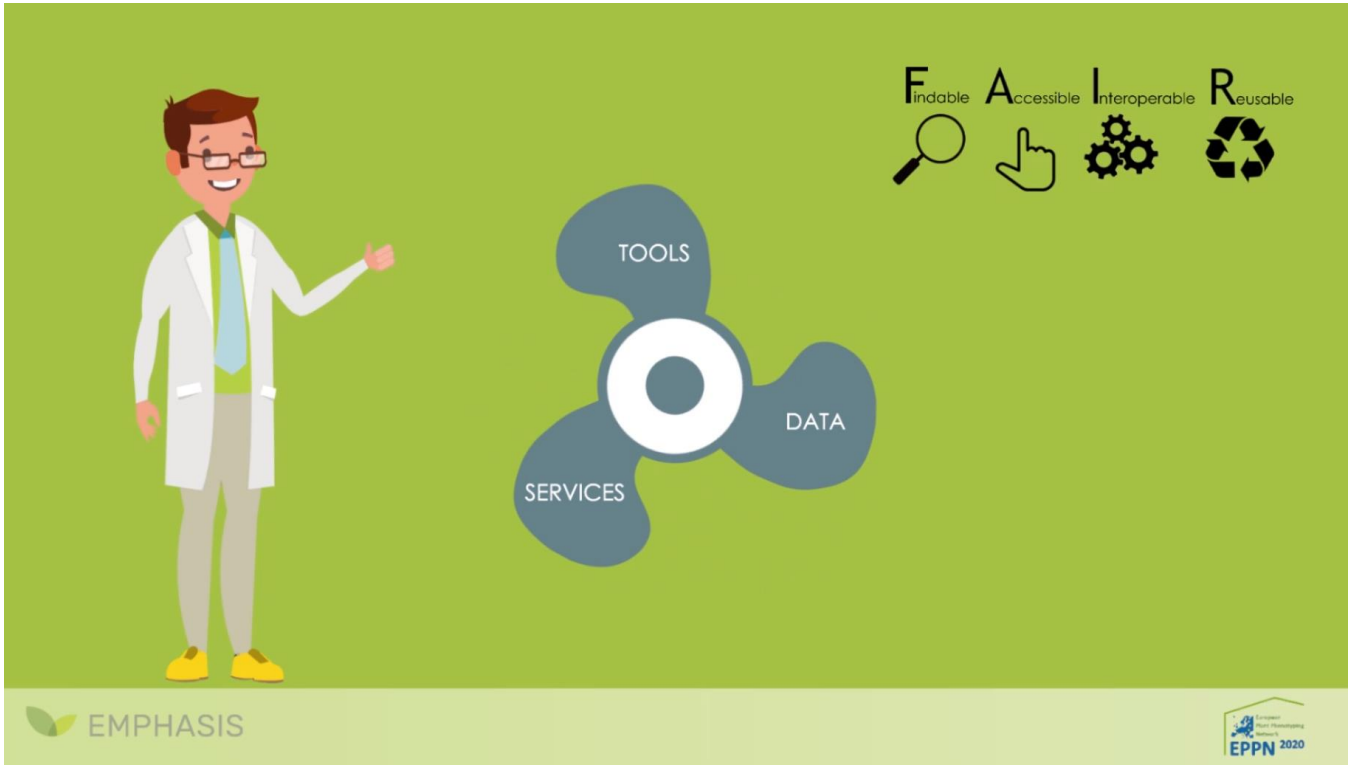
Exchange event between the European phenomic community and industry

Towards implementation of **FAIR principles** for **Plant Phenotyping**

Pascal Neveu/Isabelle Alic (INRAE, FR)

Introduction - FAIR Data – Why ?

Let’s watch a short video…
https://www.youtube.com/watch?v=rXKBy7DOPrY



Phenomics Data Challenge

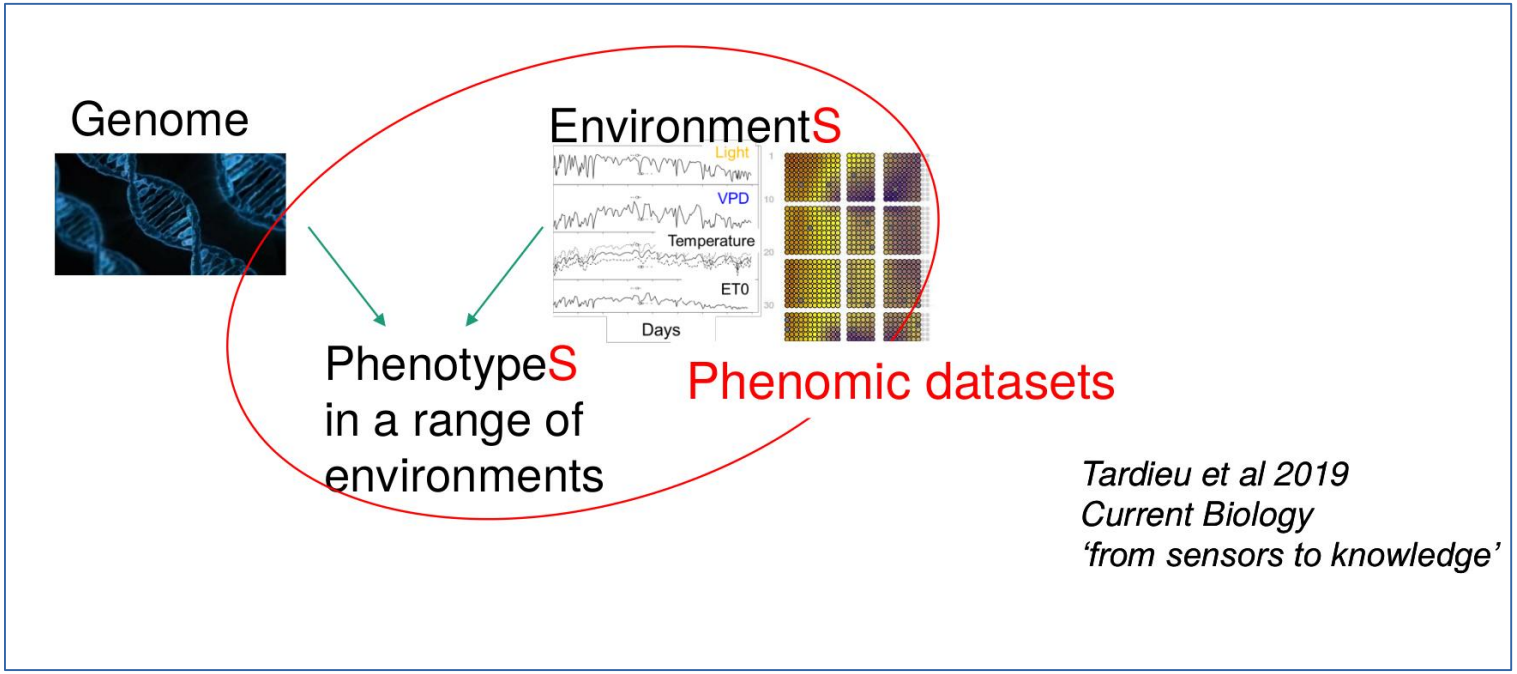
- **Plant Phenomics Experimentations**
  - Expensive, require a lot of resources and often very hard
  - Cannot be reproduced
  - Huge and complex datasets
  - Strong needs of transparence: reproducibility for data analytics

- **Save time, make data valuable!**

But re-analyses, meta-analyses and new analyses
→ impossible without rich metadata

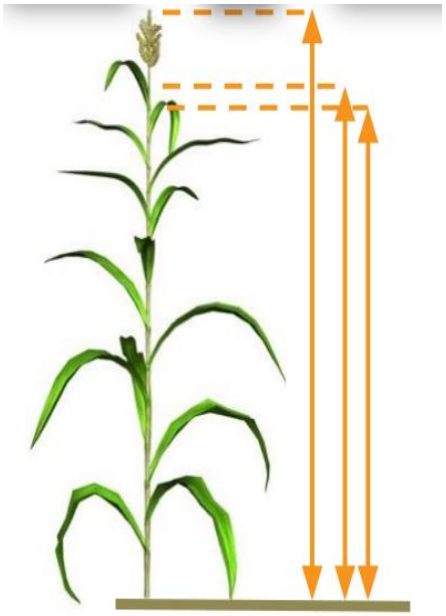
Obstacles for data reuse: Environment (time +space)

- **EMPHASIS Information systems**
  - **Environmental conditions** are PART of the phenomic information
  - We need the **time courses of environmental variables**, at the **exact time** (minutes) when data were collected
  - And the **precise x-y-z position** of sensors of measured plants/plots

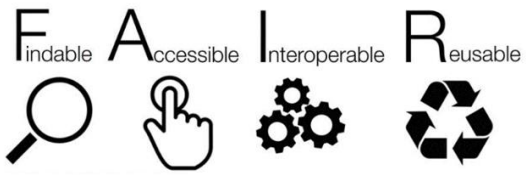


Some common mistakes we do

- **Metadata in file names** (low and very poor metadata)
- **Ambiguous ID**
- **Variable naming**
  same name for several variables, not well defined, no ID, no schema,…
- **Data are stored on personal computer**
- **Unstable data files** (non-automatisable processing)
- **Context, faults are not described**
- **No data links**
- **Missing data representation**
- **No licence**

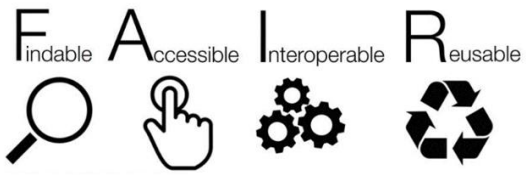


FAIR Data - Findable



Findable: Persistent ID, indexed in portals, standardized and relevant metadata

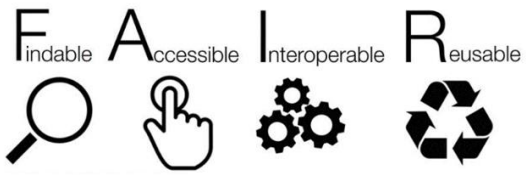
FAIR Data - Accessible



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

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

FAIR Data -Interoperable

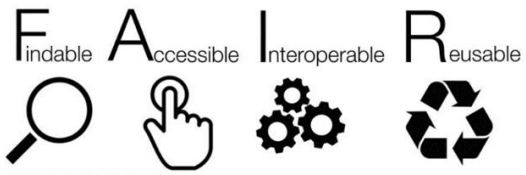


Findable: **PID**, indexed in portals, standardized and relevant metadata

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

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

FAIR Data - Reusable



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

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

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

Reusable: provenance, relevant metadata for understanding across disciplines

How ?

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)**

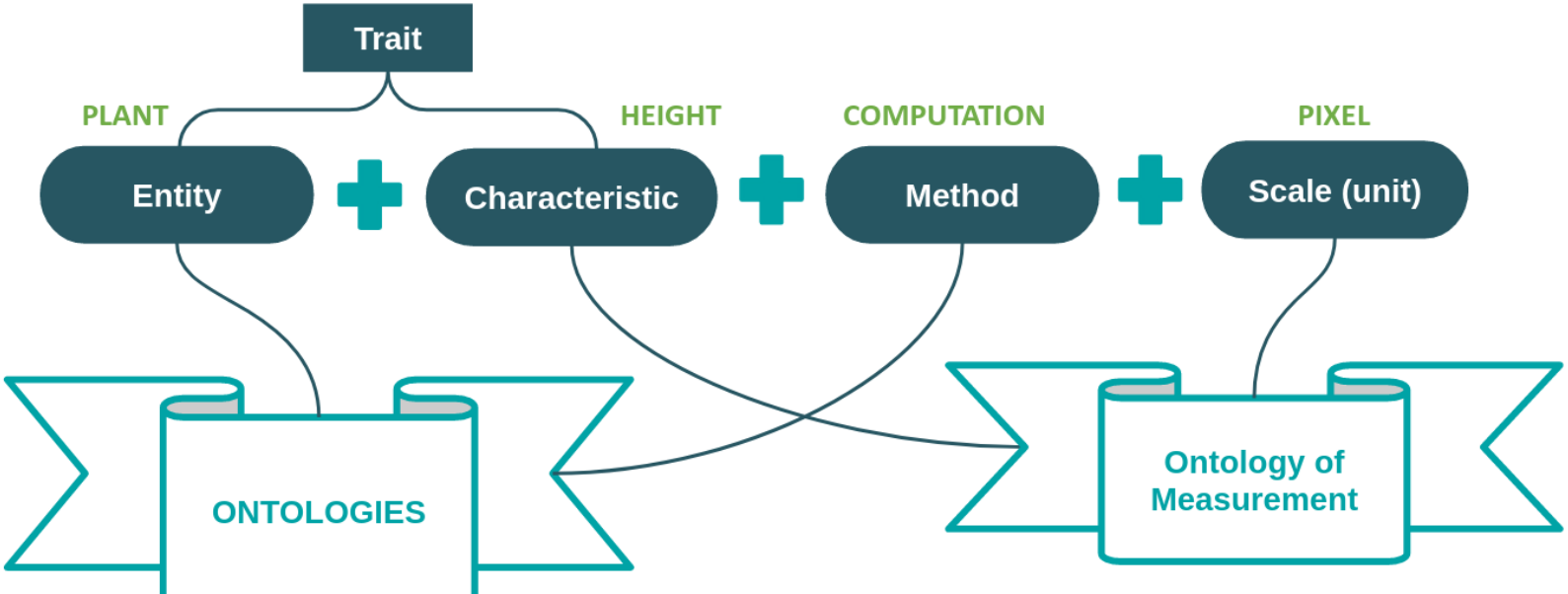
- Controlled vocabulary
- Formalized relationships between entities
- Data annotation and enrichment (search engine friendly)

Identification

URI : Uniform Resource Identifier
- **Standardized**, **Unambigous**, Actionable
- **Generated by tools** under responsability of scientific coordinator
- Use URI **for every objects**

Variable naming

- Use **URI for unambiguous name** (in global context)
- **Reuse existing variable** if available
- Use **standardized/shared representation schema** for formalisation of new variable



Identification and naming convention in **Emphasis**

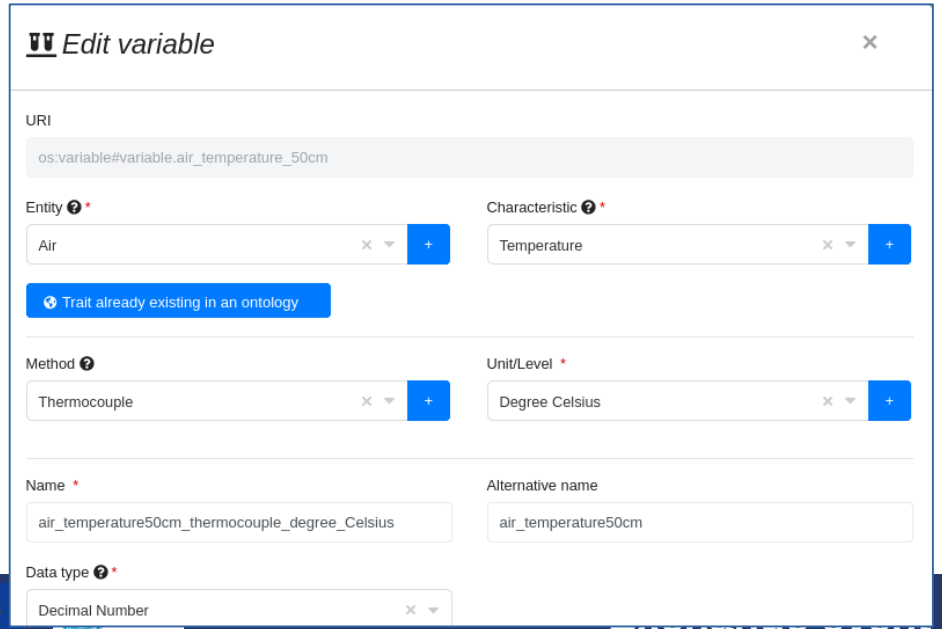
Identification

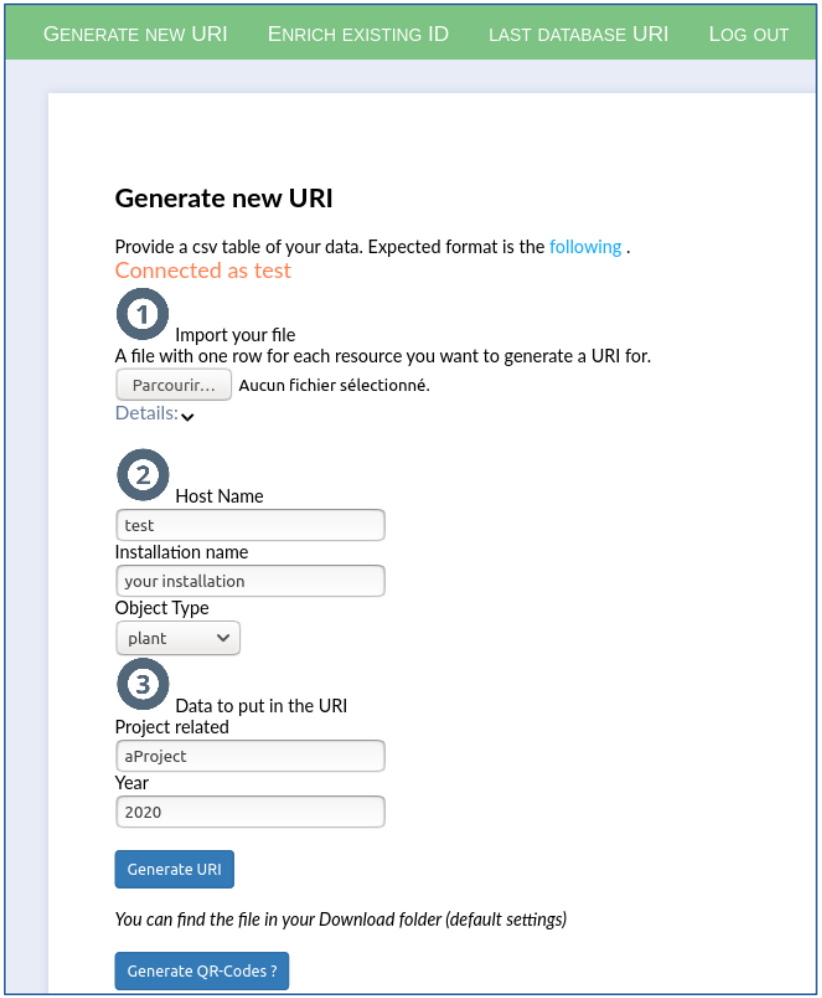
- Local infrastructure use **URIs for all objects**
- Information systems implemented in **nodes**
- Automatic generation of URIs : **ID Generator**

Variable naming

- **Environmental and phenotypic** variables named
- Variable names with **quadruple definition**

  Entity – Characteristic – Method – Unit
- Web Interface to **share and declare variables**



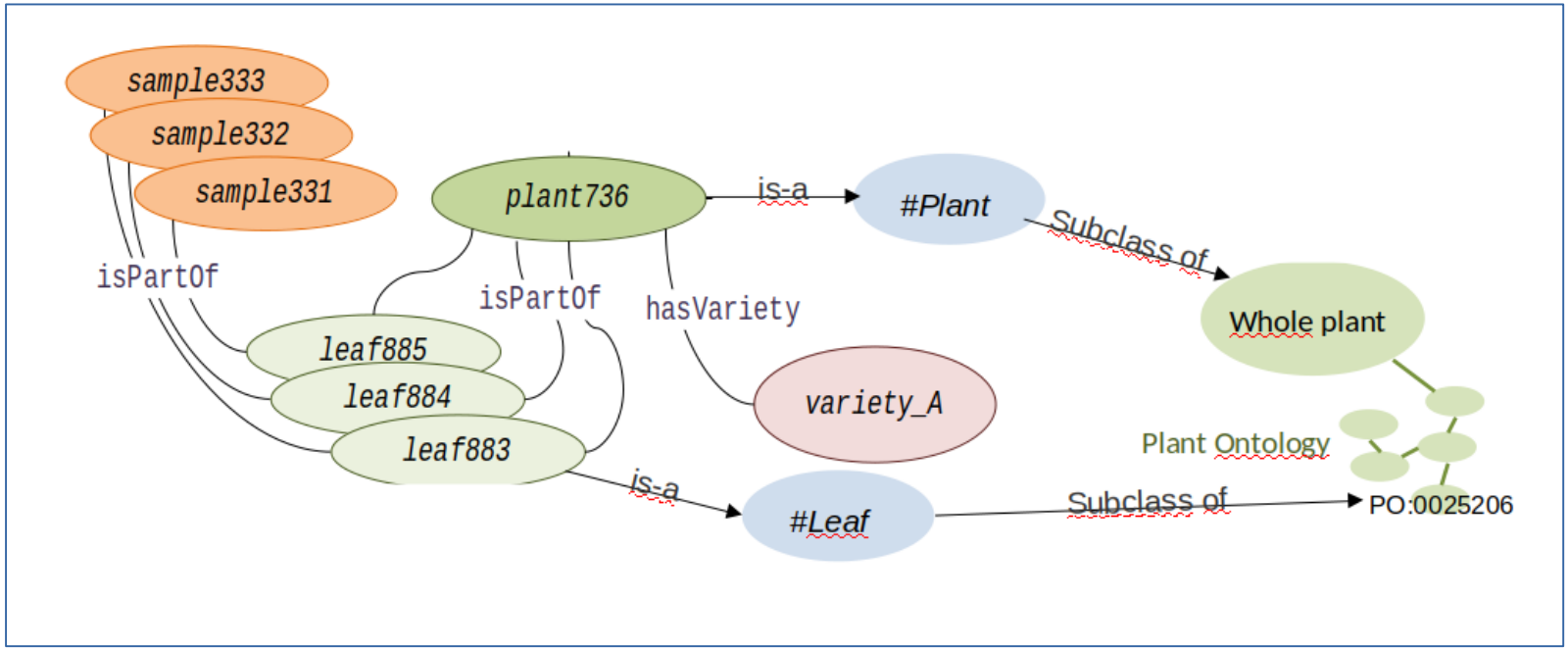


http://138.102.159.36:8082/app/

Metadata / ontologies provide the meaning of data:

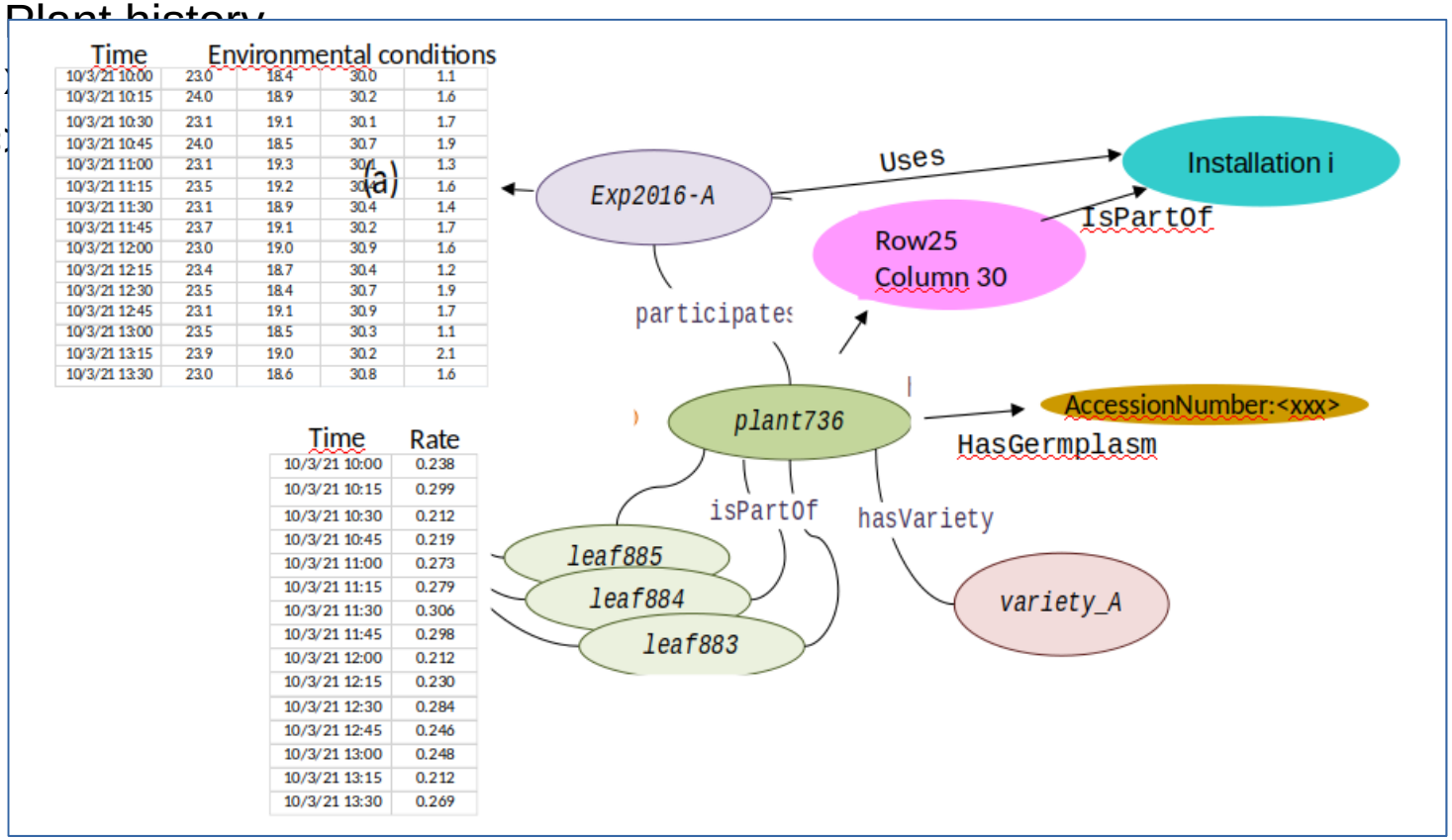
- Link each data element with a controlled, shared and machine readable vocabulary
- Structure the data in a graph

- **Reference ontologies** *(Agrovoc, Plant Ontology, PATO, etc.)*
- **Application ontologies** *(scientific objects, events, ...)*



Data are **structured by a set of ontologies**

What is kept in **EMPHASIS-compatible information systems** (e.g. PHIS) :
- Genotype,
- Accession,
- Sampling time,
- Plant history
-
=



Time | Environmental conditions

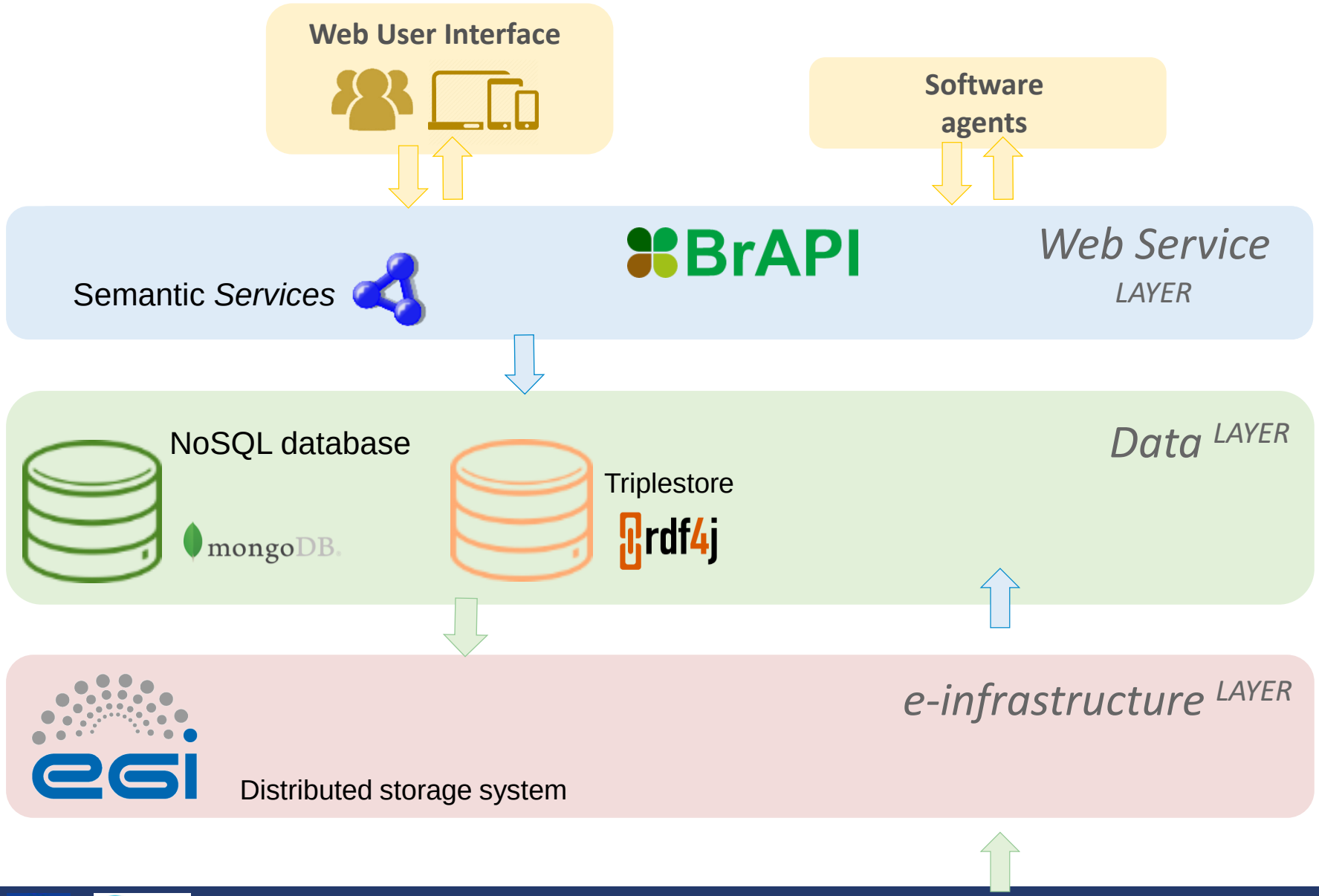
10/3/21 10:00	23.0	18.4	30.0	1.1
10/3/21 10:15	24.0	18.9	30.2	1.6
10/3/21 10:30	23.1	19.1	30.1	1.7
10/3/21 10:45	24.0	18.5	30.7	1.9
10/3/21 11:00	23.1	19.3	30.4	1.3
10/3/21 11:15	23.5	19.2	30.4	1.6
10/3/21 11:30	23.1	18.9	30.4	1.4
10/3/21 11:45	23.7	19.1	30.2	1.7
10/3/21 12:00	23.0	19.0	30.9	1.6
10/3/21 12:15	23.4	18.7	30.4	1.2
10/3/21 12:30	23.5	18.4	30.7	1.9
10/3/21 12:45	23.1	19.1	30.9	1.7
10/3/21 13:00	23.5	18.5	30.3	1.1
10/3/21 13:15	23.9	19.0	30.2	2.1
10/3/21 13:30	23.0	18.6	30.8	1.6

(a)

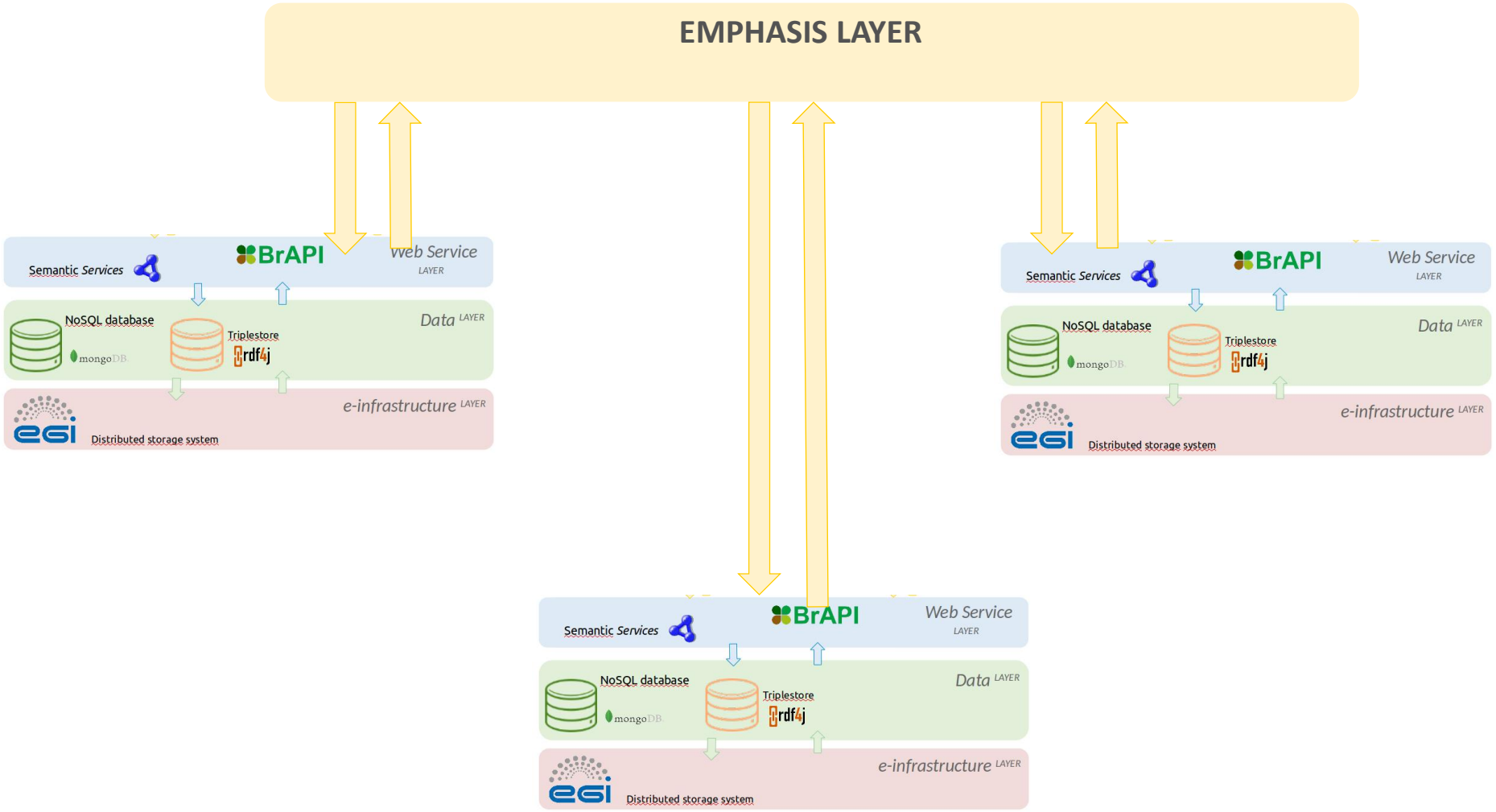
Time | Rate

10/3/21 10:00	0.238
10/3/21 10:15	0.299
10/3/21 10:30	0.212
10/3/21 10:45	0.219
10/3/21 11:00	0.273
10/3/21 11:15	0.279
10/3/21 11:30	0.306
10/3/21 11:45	0.298
10/3/21 12:00	0.212
10/3/21 12:15	0.230
10/3/21 12:30	0.284
10/3/21 12:45	0.246
10/3/21 13:00	0.248
10/3/21 13:15	0.212
10/3/21 13:30	0.269

Local Layer Architecture

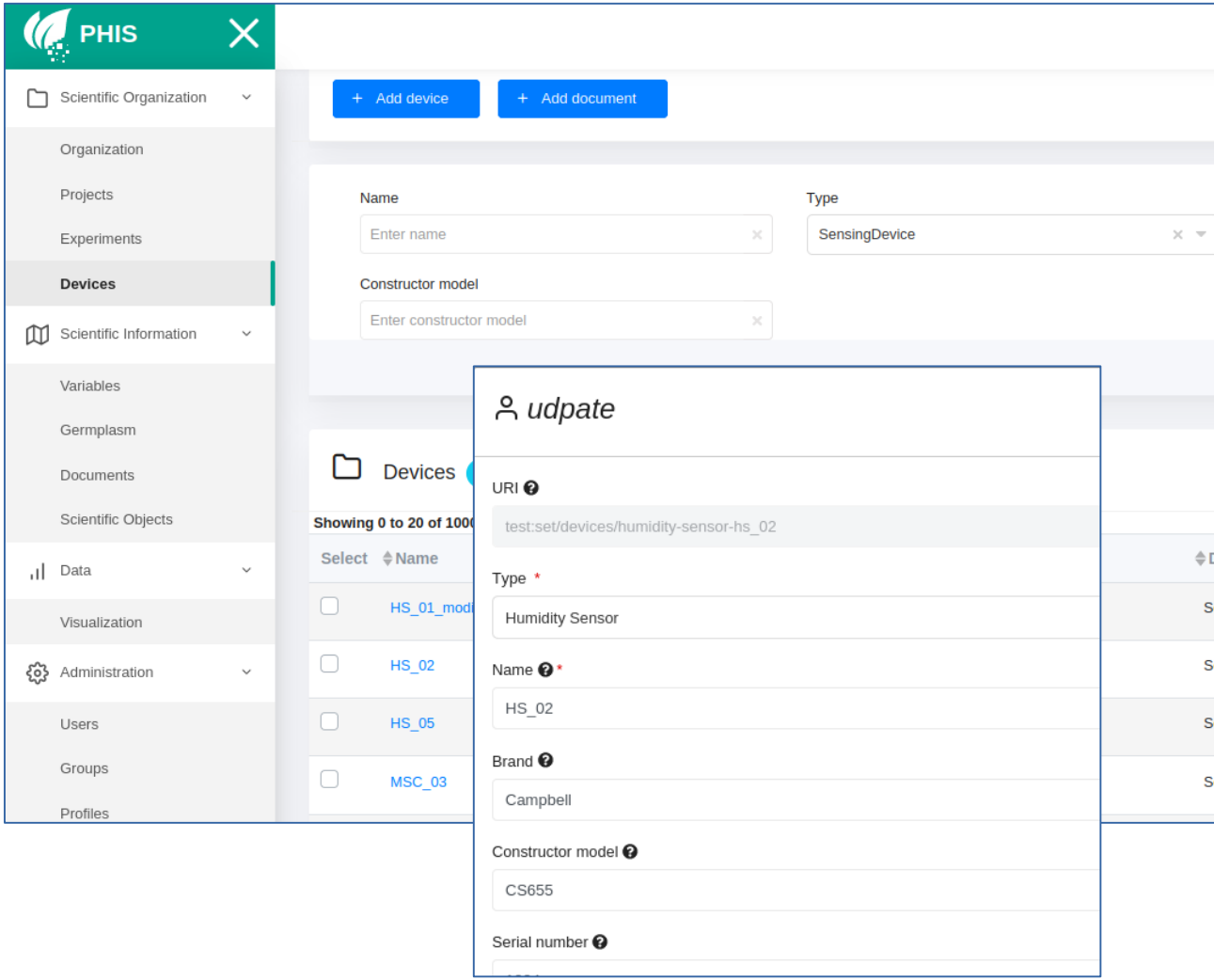


Global Layer Architecture



PHIS Interfaces allow management of

- **Scientific Organization**
  - Project information
  - Experiment
  - Facilities
  - Sensors
- **Scientific Information**
  - Variables
  - Scientific objects
  - Germplasm
  - Experimental factors
- **Data**
  - Data visualization
  - Data provenance



FAIR Data … why and how ?

✓**Data management needs time and work** at first but…

- Will **save you time** !
- Allows you to handle **large and heterogenous data sets**
- Allows you to **do analyses previously impossible**
- Allows teams and communities a **better formalization of concepts and data**

✓**FAIR data** requires **training and support**