EMPHASIS

Data Management
S1: Introduction to FAIR data

Monday 20/11/2023 - 14:00-16:00 (CET)

Isabelle Alic (INRAE)
Farzaneh Kazemipour-Ricci (INRAE)
Pascal Neveu (INRAE)

EMPHASIS

PROGRAMME

WHEN	WHAT	WHO
Session 1 Monday 20/11/2023 14:00-16:00 (CET)	INTRODUCTION TO FAIR DATA: WHY AND HOW TO MAKE FAIR DATA?	Pascal Neveu, INRAE Farzaneh Kazemipour, INRAE
Session 2 Monday 27/11/2023 14:00-16:00 (CET)	EXPERIMENTAL DATA (1) STORAGE	Pascal Neveu, INRAE
Session 3 Monday 04/12/2023 14:00-16:00 (CET)	EXPERIMENTAL DATA (2) DESCRIPTION	Farzaneh Kazemipour, INRAE Isabelle Alic, INRAE
Session 4 Thursday 07/12/2023 14:00-16:00 (CET)	ADVANCED DATA MANAGEMENT (1) VARIABLES & MEASUREMENTS	Farzaneh Kazemipour, INRAE Isabelle Alic, INRAE
Session 5 Monday 11/12/2023 14:00-16:00 (CET)	ADVANCED DATA MANAGEMENT (2) DATA MINING	Farzaneh Kazemipour. INRAE Isabelle Alic, INRAE

General objectives: Overview of data management for plant phenotyping - focus on FAIR data

Session 1

Introduction to FAIR data :

why and how to make FAIR?

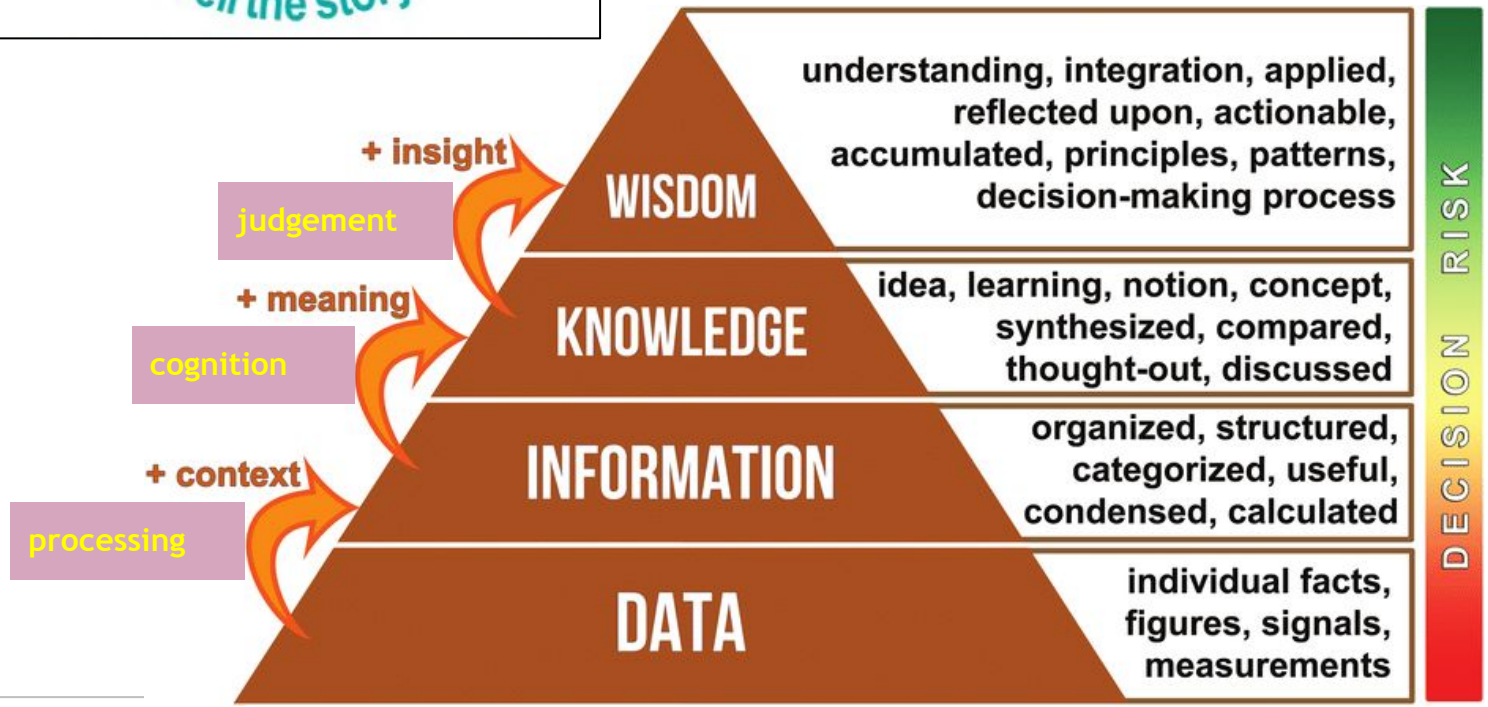
Overview

FAIR Data

- Introduction
- Data structuring
- Metadata
- Ontology & Web semantic
- Standards & APIs

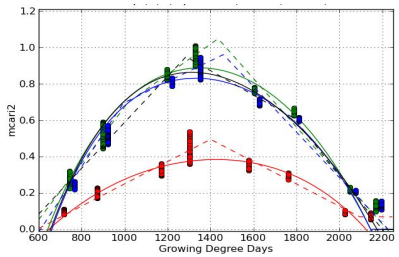
Data Value chain

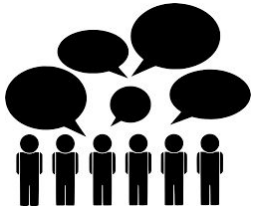
Data Pyramid

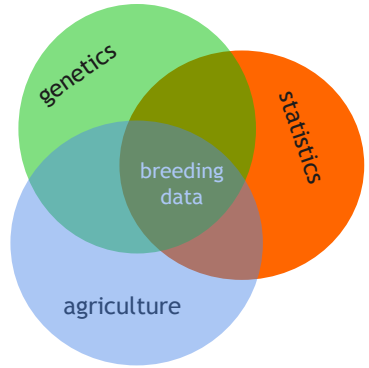


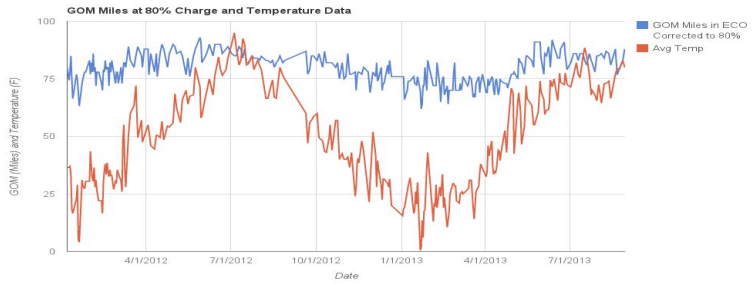
Examples of different data sources in plant sciences

- **Breeding data**
- **Crop data**
- **Weather data**
- **Soil data**
- **Environmental data**
- **Genomic data**
- **Economic, health etc.**



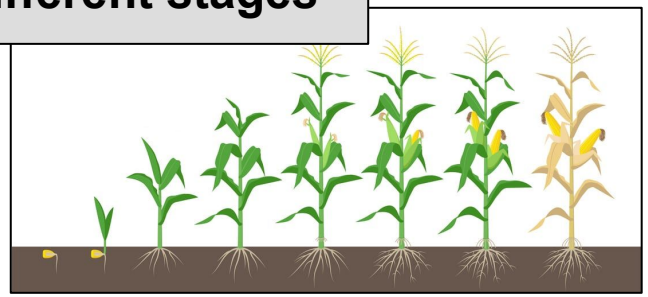






Complex Data

Different stages



Different ecosystems

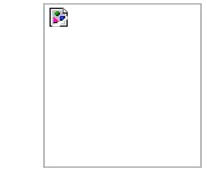
Different sources

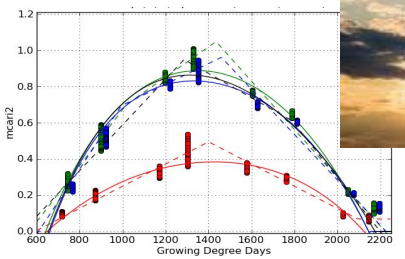
different transformations

Different scales

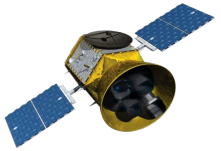
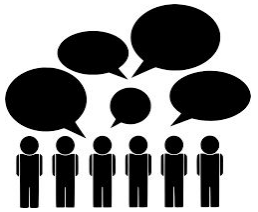
Different interactions

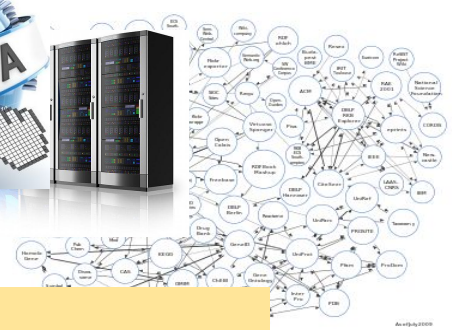
Complex Data (Source, link, description, …)





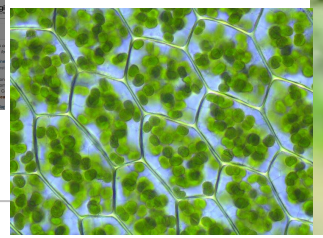


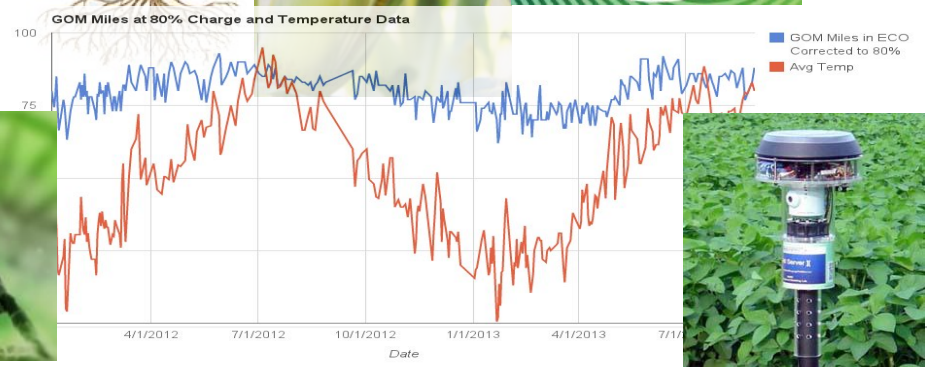







- **Orphan data →Worthless!**
- **Valuable data : if and only if structured**
- **(re)analyse, meta-analysis, visualisation, etc**

Plant Phenotyping Data sources collected by different teams

« omics » Platforms

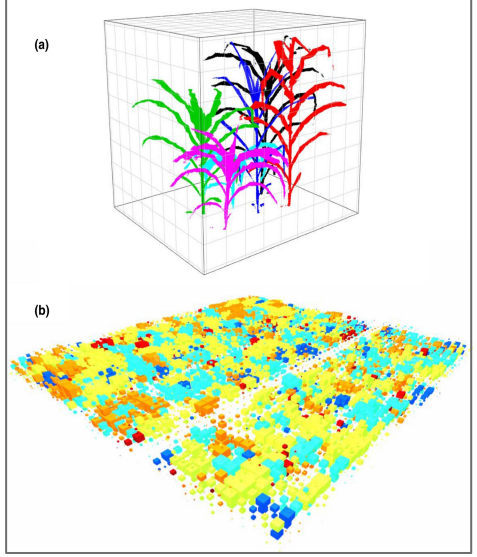
Various data complex types

Genomics

Composition and the structure of biopolymers

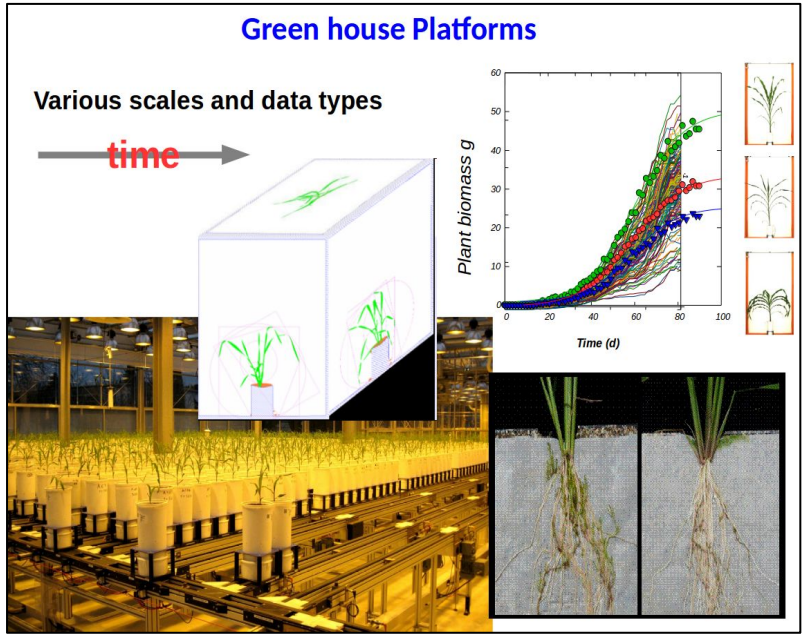
Quantification of metabolites and enzyme activities

Simulations & virtual experiments



Green house Platforms

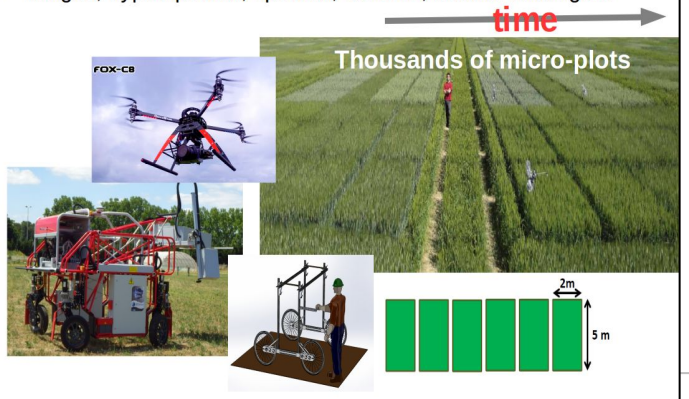
Various scales and data types



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

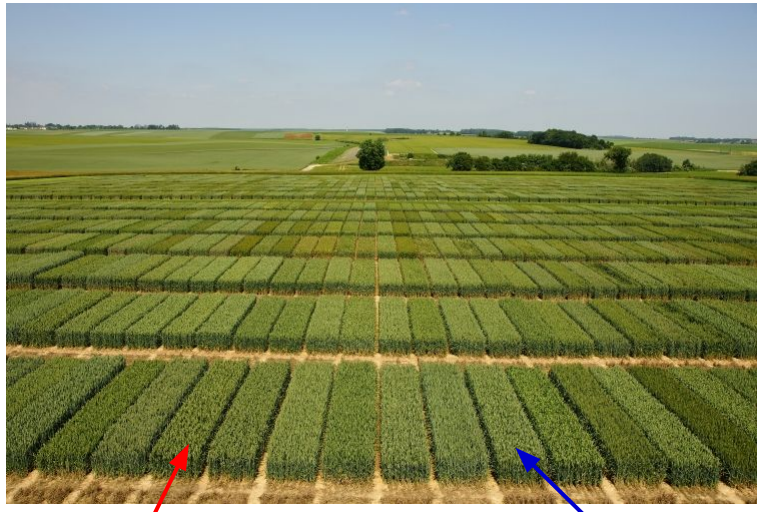
- organ, plant, population, site
- Images, sensors, human readings...

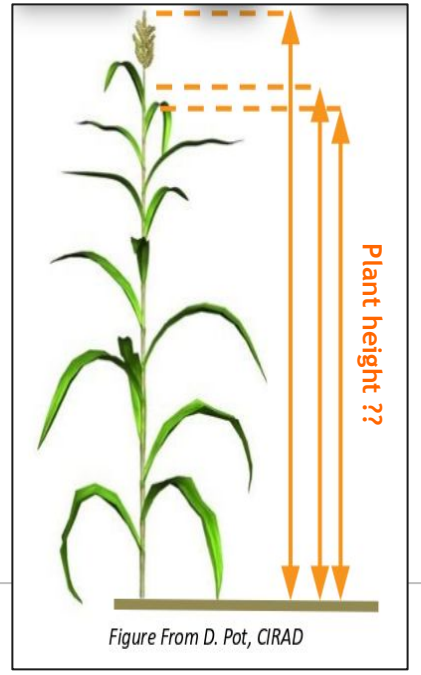
Some common mistakes we used to do

- Same name for different objects or variables (Ambiguous ID)
- Several names for same objects/variables
- Share unstable datasets
- Storage on personal computers/discs
- Lack of context and description on data (MetaData)
- Errors and missing data are not properly identified
- Use of conceptual information (not machine readable : colors, fonts, …)
- No link between data
- Data processing steps not tracked properly (softwares, parameters, models, …)
- Raw and processed data are mixed
- …

Some common mistakes we used to do

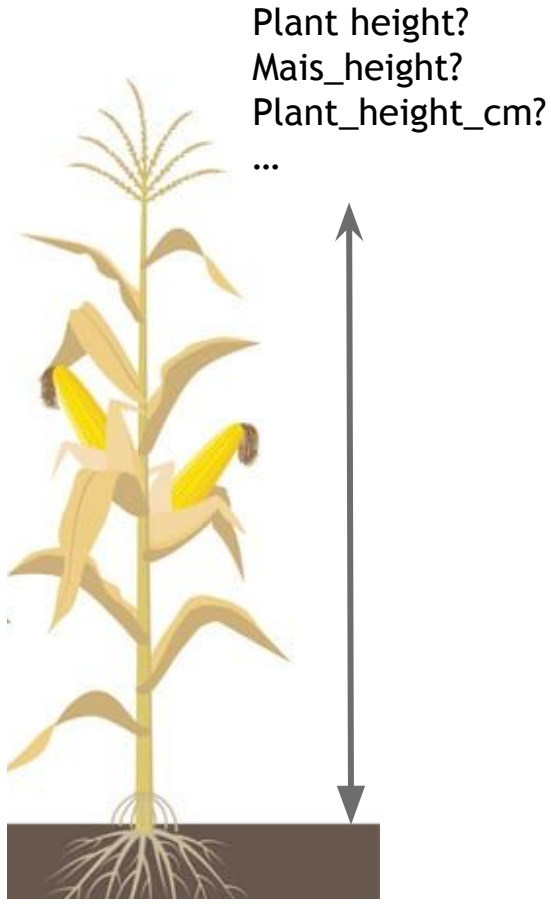
- ***Same name for different objects or variables (Ambiguous ID)***
- Several names for same objects/variables
- Share unstable datasets
- Storage on personal computers/discs
- Lack of context and description on data (MetaData)
- Errors and missing data are not properly identified
- Use of conceptual information (not machine readable : colors, fonts, …)
- No link between data
- Data processing steps not tracked properly (softwares, parameters, models, …)
- Raw and processed data are mixed
- …





Some common mistakes we used to do

- Same name for different objects or variables (Ambiguous ID)
- ***Several names for same objects/variables***
- Share unstable datasets
- Storage on personal computers/discs
- Lack of context and description on data (MetaData)
- Errors and missing data are not properly identified
- Use of conceptual information (not machine readable : colors, fonts, …)
- No link between data
- Data processing steps not tracked properly (softwares, parameters, models, …)
- Raw and processed data are mixed
- …



Some common mistakes we used to do

- Same name for different objects or variables (Ambiguous ID)
- Several names for same objects/variables
- ***Share unstable datasets***
- Storage on personal computers/discs
- Lack of context and description on data (MetaData)
- Errors and missing data are not properly identified
- Use of conceptual information (not machine readable : colors, fonts, …)
- No link between data
- Data processing steps not tracked properly (softwares, parameters, models, …)
- Raw and processed data are mixed
- …

- Customized Excel sheets & Macro
- Versioning
- …

Some common mistakes we used to do

- Same name for different objects or variables (Ambiguous ID)
- Several names for same objects/variables
- Share unstable datasets
- ***Storage on personal computers/discs***
- Lack of context and description on data (MetaData)
- Errors and missing data are not properly identified
- Use of conceptual information (not machine readable : colors, fonts, …)
- No link between data
- Data processing steps not tracked properly (softwares, parameters, models, …)
- Raw and processed data are mixed
- …

Some common mistakes we used to do

- Same name for different objects or variables (Ambiguous ID)
- Several names for same objects/variables
- Share unstable datasets
- Storage on personal computers/discs
- ***Lack of context and description on data (MetaData)***
- Errors and missing data are not properly identified
- Use of conceptual information (not machine readable : colors, fonts, …)
- No link between data
- Data processing steps not tracked properly (softwares, parameters, models, …)
- Raw and processed data are mixed
- …

- Only “Read me”
- Metadata included in the file name
- …

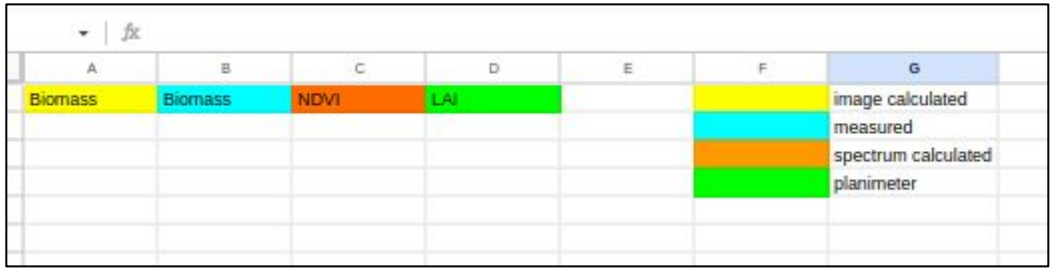
Some common mistakes we used to do

- Same name for different objects or variables (Ambiguous ID)
- Several names for same objects/variables
- Share unstable datasets
- Storage on personal computers/discs
- Lack of context and description on data (MetaData)
- ***Errors and missing data are not properly identified***
- Use of conceptual information (not machine readable : colors, fonts, …)
- No link between data
- Data processing steps not tracked properly (softwares, parameters, models, …)
- Raw and processed data are mixed
- …

- Sensor failure ⇒ ignore data
- Human errors ⇒ correct data
- 0 / NA/ “ “/
- Smoothing
- Thresholding
- …

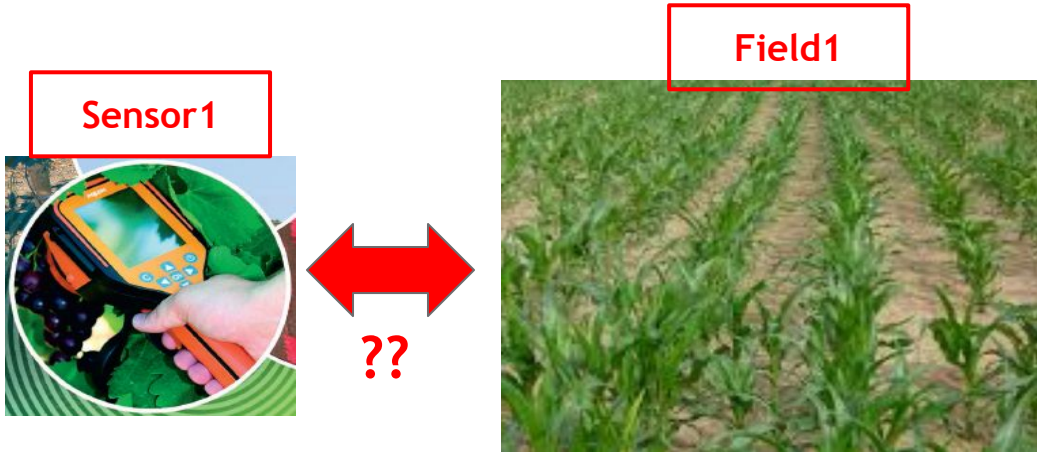
Some common mistakes we used to do

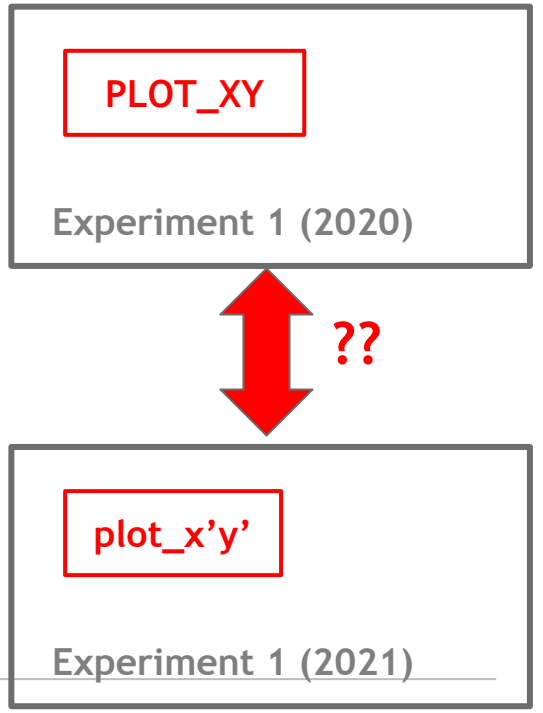
- Same name for different objects or variables (Ambiguous ID)
- Several names for same objects/variables
- Share unstable datasets
- Storage on personal computers/discs
- Lack of context and description on data (MetaData)
- Errors and missing data are not properly identified
- ***Use of conceptual information (not machine readable: colors, fonts, …)***
- No link between data
- Data processing steps not tracked properly (softwares, parameters, models, …)
- Raw and processed data are mixed
- …



Some common mistakes we used to do

- Same name for different objects or variables (Ambiguous ID)
- Several names for same objects/variables
- Share unstable datasets
- Storage on personal computers/discs
- Lack of context and description on data (MetaData)
- Errors and missing data are not properly identified
- Use of conceptual information (not machine readable : colors, fonts, …)
- ***No link between data***
- Data processing steps not tracked properly (softwares, parameters, models, …)
- Raw and processed data are mixed
- …





Some common mistakes we used to do

- Same name for different objects or variables (Ambiguous ID)
- Several names for same objects/variables
- Share unstable datasets
- Storage on personal computers/discs
- Lack of context and description on data (MetaData)
- Errors and missing data are not properly identified
- Use of conceptual information (not machine readable : colors, fonts, …)
- No link between data
- ***Data processing steps not tracked properly (softwares, parameters, models, …)***
- Raw and processed data are mixed
- …

Some common mistakes we used to do

- Same name for different objects or variables (Ambiguous ID)
- Several names for same objects/variables
- Share unstable datasets
- Storage on personal computers/discs

J	K	L	M	N	O	P
biomass (gr/m2)	LAI (image)	LAI (mean/plot)	N (%)	N (mean/plot)	Yield (simulation)	

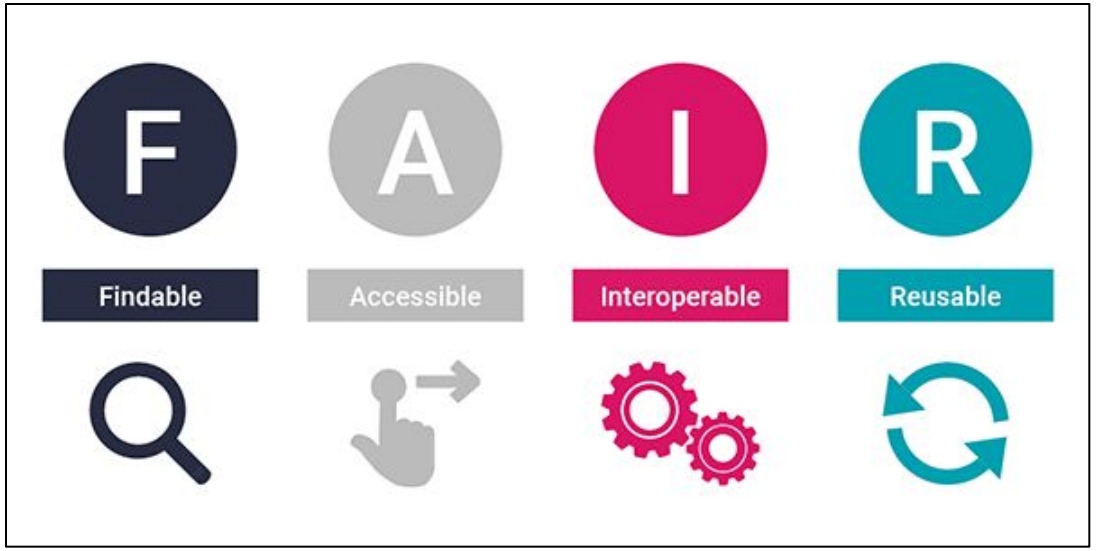
- Lack of context and description on data (MetaData)
- Errors and missing data are not properly identified
- Use of conceptual information (not machine readable : colors, fonts, …)
- No link between data
- Data processing steps not tracked properly (softwares, parameters, models, …)
- ***Raw and processed data are mixed***
- …

Have you thought about this?

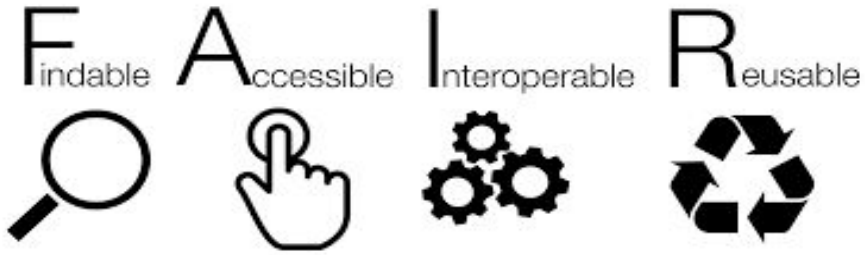
- Identification of different objects?
- Description of data?
- Private and/or sensitive data?

  Storage and security of data?
- How to share your data?

  How to access to data of other partners?
- Is there any documentation of data?
- Implementation of a Data Management Plan (DMP)
- …



Principles of FAIR Data



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

Challenge: coordinated and sustainable data services

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

Challenge: cultural vision

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

Challenge: skill development (interdisciplinary)

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

Challenge: new analysis methods

Data structuring

Data structure ⇒ store, retrieve, process data and Implement good practices:

- Make FAIR data
- Flexible
- Ability for understanding (and reproducing) data processing
- Ability to enforce DMP and Open Science

Based on two key elements:

Identification

Semantic

Identification: In a specific context, an ID must point out a unique resource

Standardized & unambiguous identification of entities:

- Studying objects (plants, plots, canopy, germplams, etc)
- Experimental context (projects, experiments, studies)
- Experimental resources (devices, facilities, vectors, etc)
- Events (management, accidents, meteo, etc.)

Semantic (based on ontology set) provides:

- **Definition (data understanding)**
- **Schema (organization) of data**
  - controlled and standardized vocabulary
  - knowledge representation models
  - Formalized relationships between entities

- Data annotation and enrichment (e.g. search engine friendly)
- A frame for reproducible data processing

Identification is crucial for leveraging data sets

URI Uniform Resource Identifier
http://www.paris.fr/monuments/tourEiffel
#tourEiffel

URL Uniform Resource Locator
https://www.w3.org/Consortium/

IRI Internationalized Resource Identifier
https://ja.wikipedia.org/wiki/ 特別:投稿記録

IRI is a superset of URI (IRI ⊃ URI)
URI is a superset of URL (URI ⊃ URL)

- Non-ambiguous
- Persistency depend on domain name
- Resolvable
- Stable

URI

- **Standardized and** easy integration in Web application
- **Unambiguous**
- **Actionable** (dereferencing)

URI → generated by tools under responsibility of local coordinator

URI of plant
<http://phenome.fr/arch/2017/c17000118>
URI of pot:
<http://phenome.fr/arch/2013/pc13001542>
URI of cart:
<http://phenome.fr/arch/2013/ct1300123>

URI of cabin:
<http://phenome.fr/arch/2018/ac180015>
URI of camera:
<http://phenome.fr/arch/2018/ac180019>

URI of image:
<m3p:arch/2017/ic17002295855>

ID allows to associate information (including MD)

http://phenome.fr/arch/2017/c17000118

LAI: 3.2
Sowing date: 12-05-2020
Experiment name: study1
Facility: Greenhouse1
Location: Montpellier
Institute: INRAE

What is MetaData

Metadata is: Data 'reporting'

- **WHO** created the data?
- **WHAT** is the content of the data?
- **WHEN** were the data created?
- **WHERE** is it geographically?
- **HOW** were the data developed?
- **WHY** were the data developed?

What is Metadata

DataONE

What is MetaData

The 6 levels of metadata management

Metadata Level 1: Description

Human readable

description2014Bp45.txt

Id : http://www.inrae.fr/PechRouche/2014Bp45
Plot Beausoleil
Site = Pech Rouge
Position = « R4-P5 »
Carignan
The plot is supervising by Jean
planted : 2014

description2013Vp56.txt

ID : http://www.inrae.fr/PechRouche/2013Vp56
Verson Sud
Farm = Pech-Rouge
Location = [5,6]
Carignan
The plot manager Jean Dupont
planted : 2013

Metadata Level 1: Description

Semantic resources: none

Tools: editor, word processor, etc.

description2014Bp45.txt

Id : http://www.inrae.fr/PechRouche/2014Bp45
Plot Beausoleil
Site = Pech Rouge
Position = « R4-P5 »
Carignan
The plot is supervising by Jean
planted : 2014

description2013Vp56.txt

ID : http://www.inrae.fr/PechRouche/2013Vp56
Verson Sud
Farm = Pech-Rouge
Location = [5,6]
Carignan
The plot manager Jean Dupont
planted : 2013

Metadata Level 2: Description + **Syntax**

Machine accessible

key/value approach

description2014Bp45.jsn

```
”ID” : ”http://www.inrae.fr/PechRouche/2014Bp45”
”Plot” :  ”Beausoleil”
”Site” : ”Pech Rouge”
”Position” : ”R4-P5”
”variety” : ”carignan”
”supervisor” : ”Jean”
”planted-year” : ”2014”
```

description2013Vp56.jsn

```
”ID” : ”http://www.inrae.fr/PechRouche/2013Vp56”
”plot” : ”Verson Sud”
”farm” : ”Pech-Rouge”
”Location” : ”[5,6]”
”Variety” : ”Carignan”
”Manager” : ”Jean Dupont”
”planted” : ”2013”
```

Metadata Level 2: Description + **Syntax**

Semantic resources: none

Tools: information representation languages such as XML or JSON, spreadsheet

description2014Bp45.jsn

```
”ID” : ”http://www.inrae.fr/PechRouche/2014Bp45”
”Plot” :  ”Beausoleil”
”Site” : ”Pech Rouge”
”Position” : ”R4-P5”
”variety” : ”carignan”
”supervisor” : ”Jean”
”planted-year” : ”2014”
```

description2013Vp56.jsn

```
”ID” : ”http://www.inrae.fr/PechRouche/2013Vp56”
”plot” : ”Verson Sud”
”farm” : ”Pech-Rouge”
”Location” : ”[5,6]”
”Variety” : ”Carignan”
”Manager” : ”Jean Dupont”
”planted” : ”2013”
```

Metadata Level 3: Description + Syntax + **Vocabulary**

Machine readable

description2014Bp45.jsn

```
”ID” : ”http://www.inrae.fr/PechRouche/2014Bp45”
”plot” :  ”Beausoleil”
”farm” : ”Pech Rouge”
”location” : ”R4-P5”
”variety” : ”carignan”
”supervisor” : ”Jean”
”planted-year” : ”2014”
```

description2013Vp56.jsn

```
”ID” : ”http://www.inrae.fr/PechRouche/2013Vp56”
”plot” : ”Verson Sud”
”farm” : ”Pech-Rouge”
”location” : ”[5,6]”
”variety” : ”Carignan”
”supervisor” : ”Jean Dupont”
”planted-year” : ”2013”
```

Metadata Level 3: Description + Syntax + **Vocabulary**

Semantic resources: standard term sets (e.g. Dublin Core), dictionary, thesaurus, formal naming from ontologies

Tools: information representation languages, spreadsheet, database systems

description2014Bp45.jsn

```
”ID” : ”http://www.inrae.fr/PechRouche/2014Bp45”
”plot” :  ”Beausoleil”
”farm” : ”Pech Rouge”
”location” : ”R4-P5”
”variety” : ”carignan”
”supervisor” : ”Jean”
”planted-year” : ”2014”
```

description2013Vp56.jsn

```
”ID” : ”http://www.inrae.fr/PechRouche/2013Vp56”
”plot” : ”Verson Sud”
”farm” : ”Pech-Rouge”
”location” : ”[5,6]”
”variety” : ”Carignan”
”supervisor” : ”Jean Dupont”
”planted-year” : ”2013”
```

Metadata Level 3: Description + Syntax + **Vocabulary**

Machine readable

description2014Bp45.jsn

```
”ID” : ”http://www.inrae.fr/PechRouche/2014Bp45”
”plot” :  ”Beausoleil”
”farm” : ”Pech Rouge”
”location” : ”R4-P5”
”variety” : ”carignan”
”supervisor” : ”Jean”
”planted-year” : ”2014”
```

description2013Vp56.jsn

```
”ID” : ”http://www.inrae.fr/PechRouche/2013Vp56”
”plot” : ”Verson Sud”
”farm” : ”Pech-Rouge”
”location” : ”[5,6]”
”variety” : ”Carignan”
”supervisor” : ”Jean Dupont”
”planted-year” : ”2013”
```

Metadata Level 4: Description + Syntax + Vocabulary + **References**

Machine readable & browsable

description2014Bp45.jsn

```
”ID” : ”http://www.inrae.fr/PechRouche/2014Bp45”
”plot” :  ”Beausoleil”
”farm” :  ”http://www.inrae.fr/PechRouge”
”location” : ”R4-P5”
”variety” :”http://www.agrisource.org/Vine/carignan”
”supervisor” : ”http://www.inrae.fr/Jean.Dupont”
”planted-year” : ”2014”
```

description2013Vp56.jsn

```
”ID” : ”http://www.inrae.fr/PechRouche/2013Vp56”
”plot” : ”Verson Sud”
”farm” : ”http://www.inrae.fr/PechRouge”
”location” : ”[5,6]”
”variety” : ”http://www.agrisource.org/Vine/carignan”
”supervisor” : ”Jean Dupont”
”planted-year” : ”2013”
```

Metadata Level 4: Description + Syntax + Vocabulary + **References**

Semantic resources: standard term sets, thesaurus, dictionary, formal naming from ontologies

Tools: browsers + information representation languages, spreadsheet, database systems

description2014Bp45.jsn

```
”ID” : ”http://www.inrae.fr/PechRouche/2014Bp45”
”plot” :  ”Beausoleil”
”farm” :  ”http://www.inrae.fr/PechRouge”
”location” : ”R4-P5”
”variety” :”http://www.agrisource.org/Vine/carignan”
”supervisor” : ”http://www.inrae.fr/Jean.Dupont”
”planted-year” : ”2014”
```

description2013Vp56.jsn

```
”ID” : ”http://www.inrae.fr/PechRouche/2013Vp56”
”plot” : ”Verson Sud”
”farm” : ”http://www.inrae.fr/PechRouge”
”location” : ”[5,6]”
”variety” : ”http://www.agrisource.org/Vine/carignan”
”supervisor” : ”Jean Dupont”
”planted-year” : ”2013”
```

Metadata Level 4: Description + Syntax + Vocabulary + **References**

Machine readable & browsable

description2014Bp45.jsn

```
”ID” : ”http://www.inrae.fr/PechRouche/2014Bp45”
”plot” : ”http://www.inrae.fr/PechRouche/Beausoleil”
”farm” : ”http://www.inrae.fr/PechRouge”
”location” : ”R4-P5”
”variety” :”http://www.agrisource.org/Vine/carignan”
”supervisor” : ”http://www.inrae.fr/Jean.Dupont”
”planted-year” : ”2014”
```

description2013Vp56.jsn

```
”ID” : ”http://www.inrae.fr/PechRouche/2013Vp56”
”plot” : ”http://www.inrae.fr/PechRouche/VersonSud”
”farm” : ”http://www.inrae.fr/PechRouge”
”location” : ”[5,6]”
”variety” : ”http://www.agrisource.org/Vine/carignan”
”supervisor” : ”Jean Dupont”
”planted-year” : ”2013”
```

Metadata Level 5: Description + Syntax + Vocabulary + **Link**

farm

Machine interpretable
Reference is managed as link between
typed objects

person

variety

plot

plot

plant

plant

Metadata Level 5: Description + Syntax + Vocabulary + **Link**

farm ”http://www.inrae.fr/PechRouge”

Machine interpretable
Reference is managed as link between
typed objects

person ”http://www.inrae.fr/Jean.Dupont”

variety ”http://www.agrisource.org/Vine/carignan”

plot ”http://www.inrae.fr/PechRouche/Beausoleil”
”planted-year” : ”2014”

plot ”http://www.inrae.fr/PechRouche/VersonSud
”planted-year” : ”2013”

plant ”http://www.inrae.fr/PechRouche/2014Bp45”
”location” : ”[4,5]”

plant ”http://www.inrae.fr/PechRouche/2013Vp56”
”location” : ”[5,6]”

Metadata Level 5: Description + Syntax + Vocabulary + **Link**

Machine interpretable
Reference is managed as link between **typed objects**

farm ”http://www.inrae.fr/PechRouge”

supervisor

person ”http://www.inrae.fr/Jean.Dupont”

part-of

part-of

variety ”http://www.agrisource.org/Vine/carignan”

has_variety

has_variety

plot ”http://www.inrae.fr/PechRouche/Beausoleil”
”planted-year” : ”2014”

plot ”http://www.inrae.fr/PechRouche/VersonSud
”planted-year” : ”2013”

is_located

is_located

plant ”http://www.inrae.fr/PechRouche/2014Bp45”
”location” : ”[4,5]”

plant ”http://www.inrae.fr/PechRouche/2013Vp56”
”location” : ”[5,6]”

Metadata Level 5: Description + Syntax + Vocabulary + **Link**

Machine interpretable

as link between
cts

farm "http://ww

Semantic resources: standard term sets, thesaurus, ontologies

Tools: RDF, JSON-LD, graph database systems

.fr/Jean.Dupont"

/carignan"

part-of

has_variety

plot "http://www.inrae.fr/PechRouche/Beausoleil"
"planted-year" : "2014"

plot "http://www.inrae.fr/PechRouche/VersonSud
"planted-year" : "2013"

is_located

is_located

plant "http://www.inrae.fr/PechRouche/2014Bp45"
"location" : "[4,5]"

plant "http://www.inrae.fr/PechRouche/2013Vp56"
"location" : "[5,6]"

Metadata Level 6: Description + Syntax + Vocabulary + **Link + Inference**

Machine interpretable
Reference is managed as link between
typed objects

farm ”http://www.inrae.fr/PechRouge”

supervisor

person ”http://www.inrae.fr/Jean.Dupont”

part-of

part-of

variety ”http://www.agrisource.org/Vine/carignan”

variety

variety

plot ”http://www.inrae.fr/PechRouche/Beausoleil”
”planted-year” : ”2014”

plot ”http://www.inrae.fr/PechRouche/VersonSud
”planted-year” : ”2013”

located

located

plant ”http://www.inrae.fr/PechRouche/2014Bp45”
”location” : ”[4,5]”

plant ”http://www.inrae.fr/PechRouche/2013Vp56”
”location” : ”[5,6]”

Metadata Level 6: Description + Syntax + Vocabulary + **Link + Inference**

Machine interpretable
Link between objects can be use for inference

farm ”http://www.inrae.fr/PechRouge”

supervisor

person ”http://www.inrae.fr/Jean.Dupont”

part-of

part-of

variety ”http://www.agrisource.org/Vine/carignan”

variety

variety

plot ”http://www.inrae.fr/PechRouche/Beausoleil”
”planted-year” : ”2014”

plot ”http://www.inrae.fr/PechRouche/VersonSud
”planted-year” : ”2013”

located

located

plant ”http://www.inrae.fr/PechRouche/2014Bp45”
”location” : ”[4,5]”

plant ”http://www.inrae.fr/PechRouche/2013Vp56”
”location” : ”[5,6]”

Metadata Level 6: Description + Syntax + Vocabulary + **Link + Inference**

Machine interpretable
Link between objects can be use for inference

farm ”http://www.inrae.fr/PechRouge”

supervi

person ”http://www.inrae.fr/Jean.Dupont”

part-of

part-of

variety ”http://www.agrisource.org/Vine/carignan”

variety

variety

plot ”http://www.inrae.fr/PechRouche/Beausoleil”
”planted-year” : ”2014”

plot ”http://www.inrae.fr/PechRouche/VersonSud
”planted-year” : ”2013”

located

located

plant ”http://www.inrae.fr/PechRouche/2014Bp45”
”location” : ”[4,5]”

plant ”http://www.inrae.fr/PechRouche/2013Vp56”
”location” : ”[5,6]”

Metadata Level 6: Description + Syntax + Vocabulary + **Link + Inference**

Machine interpretable

use for inference

farm "http://ww

Semantic resources: standard term sets, thesaurus, ontologies, rules

Tools: RDF, OWL, SWRL, graph database systems, reasons

Metadata → Linked Data

fr/Jean.Dupont"

carignan"

part-of

variety

plot "http://www.i
"planted-year" : "2014"

rae.fr/PechRouche/VersonSud
"planted-year" : "2013"

located

located

plant "http://www.inrae.fr/PechRouche/2014Bp45"
"location" : "[4,5]"

plant "http://www.inrae.fr/PechRouche/2013Vp56"
"location" : "[5,6]"

Metadata Level 6: Description + Syntax + Vocabulary + **Link + Inference**

Machine interpretable
Link between objects can be use for inference

farm ”http://www.inrae.fr/PechRouge”

supervi

person ”http://www.inrae.fr/Jean.Dupont”

part-of

part-of

variety ”http://www.agrisource.org/Vine/carignan”

variety

variety

plot ”http://www.inrae.fr/PechRouche/Beausoleil”
”planted-year” : ”2014”

plot ”http://www.inrae.fr/PechRouche/VersonSud
”planted-year” : ”2013”

located

located

plant ”http://www.inrae.fr/PechRouche/2014Bp45”
”location” : ”[4,5]”

plant ”http://www.inrae.fr/PechRouche/2013Vp56”
”location” : ”[5,6]”

MetaData Categories

- Understanding: definition, semantic, purpose, protocol, etc
- Provenance: origine, creator, time and date, quality, modification, etc
- Administrative: owner, licence, copyrights, access rules and protocols, contacts, etc
- Structure: format, schema, size, links, etc
- Management: release, description for discovery and identification, availability

MetaData Categories

XML or ***JSON*** languages allow the representation of arbitrary data structure

★ Allow modeling, encoding, transmitting and validating information
★ Make easier information quality (description, integrity, origin, etc.)
★ Make easier interoperability
★ Make easier evolution

⇒ Many available tools

XML Illustration

Principles:

Define tags (vocabulary) for naming information elements

```
<description> this element is a bla bla bla </description>
```

Information elements must be nested (no overlap between tags)

```
<all> <part>first</part> <part>second</part></all>
```

Information elements can have attributes (properties)

```
<all lang=”en”> <part>first</part> <part>second</part></all>
```

MetaData

XML

```
68
17
```

XML

```
<humidity>68</humidity>
<temperature>17</temperature>
```

XML

```
<humidity>68</humidity>
<temperature unit="C">17</temperature>
```

XML

```
<mesures site="http://inrae.fr/34/Mauguio" date="11/10/2022">
    <humidity>68</humidity>
    <temperature unit="C">17</temperature>
</mesures>
```

XML vs JSON

XML

```
<mesures site="http://inrae.fr/34/Mauguio" date="11/10/2022">
        <humidity>68</humidity>
        <temperature unit="C">17</temperature>
    </mesures>
```

JSON

```
{ "mesures": {
        "site": "http://inrae.fr/34/Mauguio",
        "date": "11/10/2022",
        "humidity": {
            "value": 68 },
        "temperature": {
            "unit": "C"
            "value":  17 }
              }
    }
```

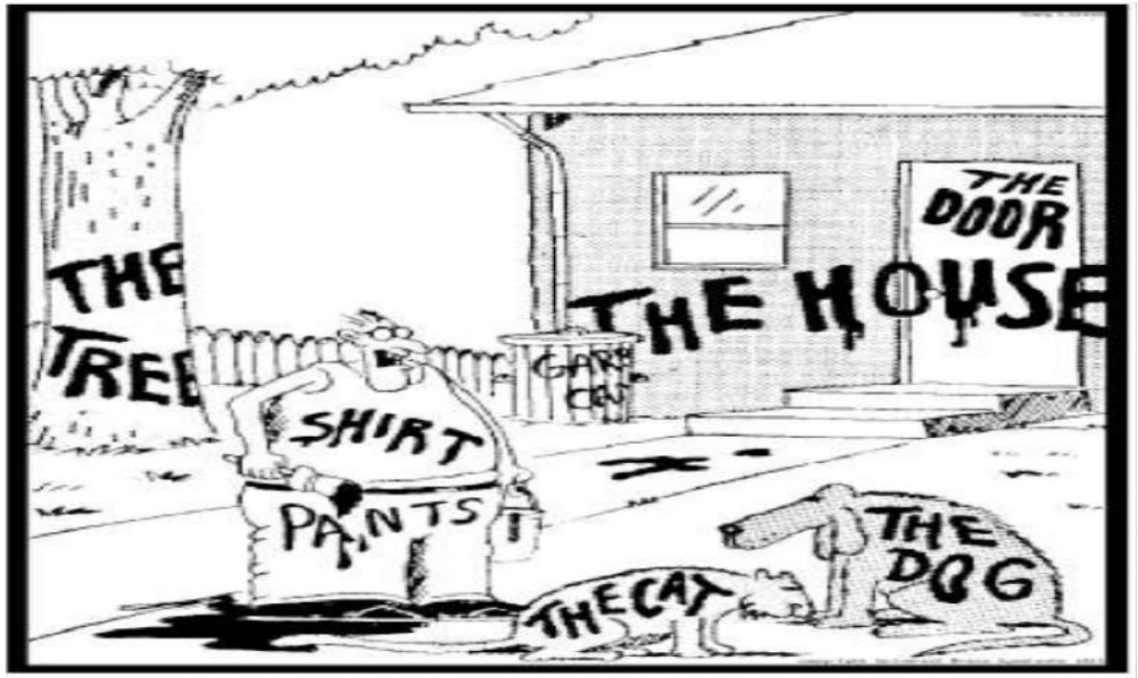
Which approach to provide MetaData?

- **Approaches :**
  - **Table/List of keywords**
  - **Tree/container**
  - **Semantic graph (RDF triple)**

Semantic Web : Common framework to share and reuse data

⇒ Implementation and use of ontologies (OWL)

Ontology: Gives meaning to data ⇒ link each element of data to a controlled and shared vocabulary



"Now! *That* should clear up a few things around here!"

A model for data description

- **R**esource **D**escription **F**ramework

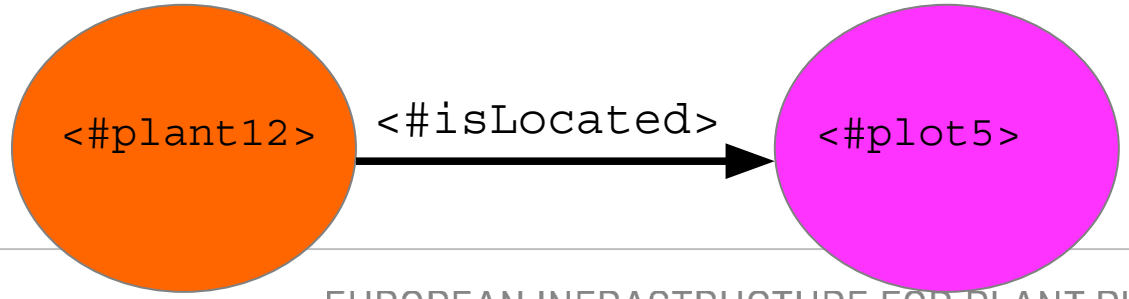
  Triple is the basic element
  (**subject**, **predicate**, **object**)

A model for data description : RDF model

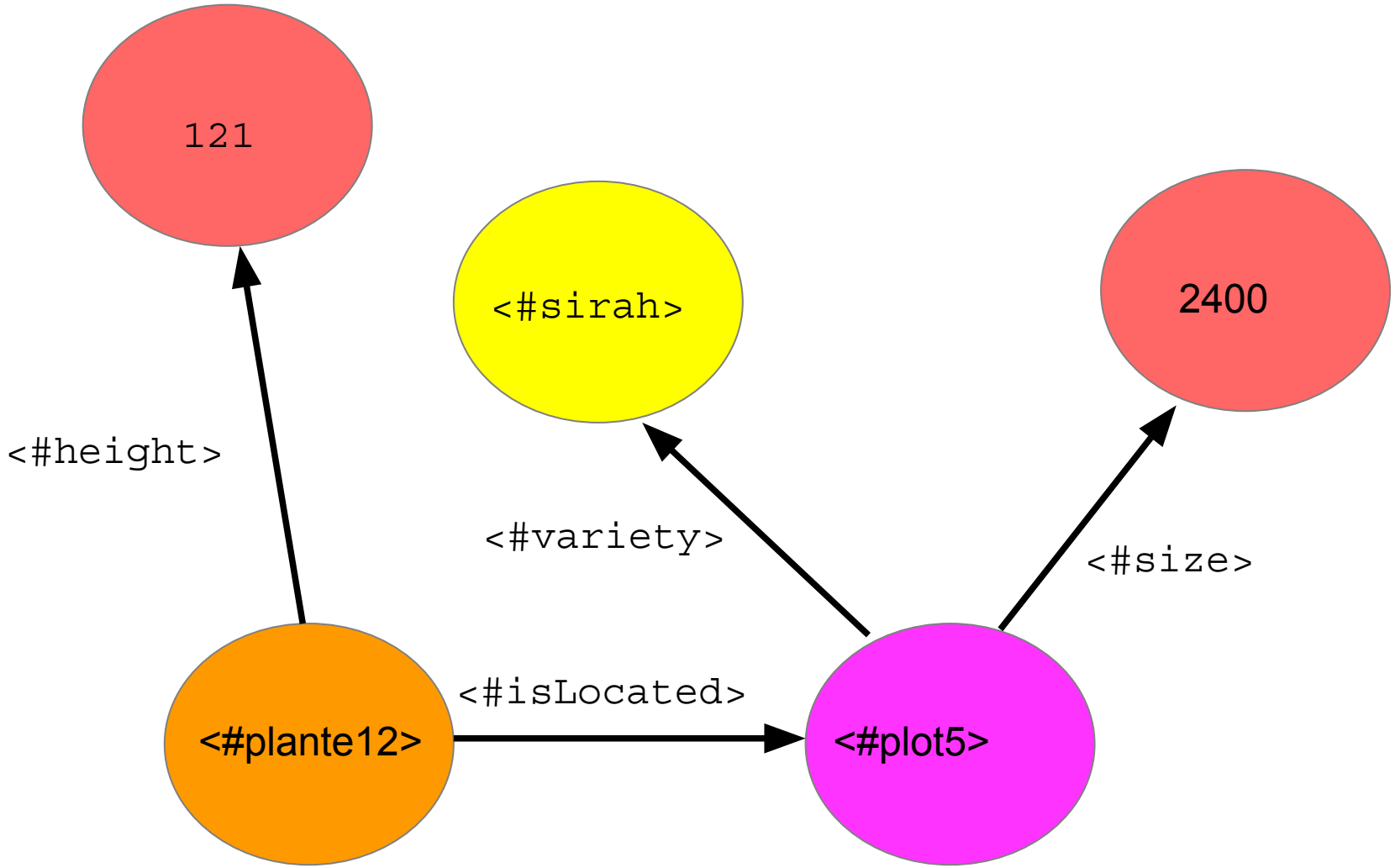
.Resource Description Framework

Triple is the basic element
(**subject**, **predicate**, **object**)

<#plant12> <#isLocated> <#plot5> .

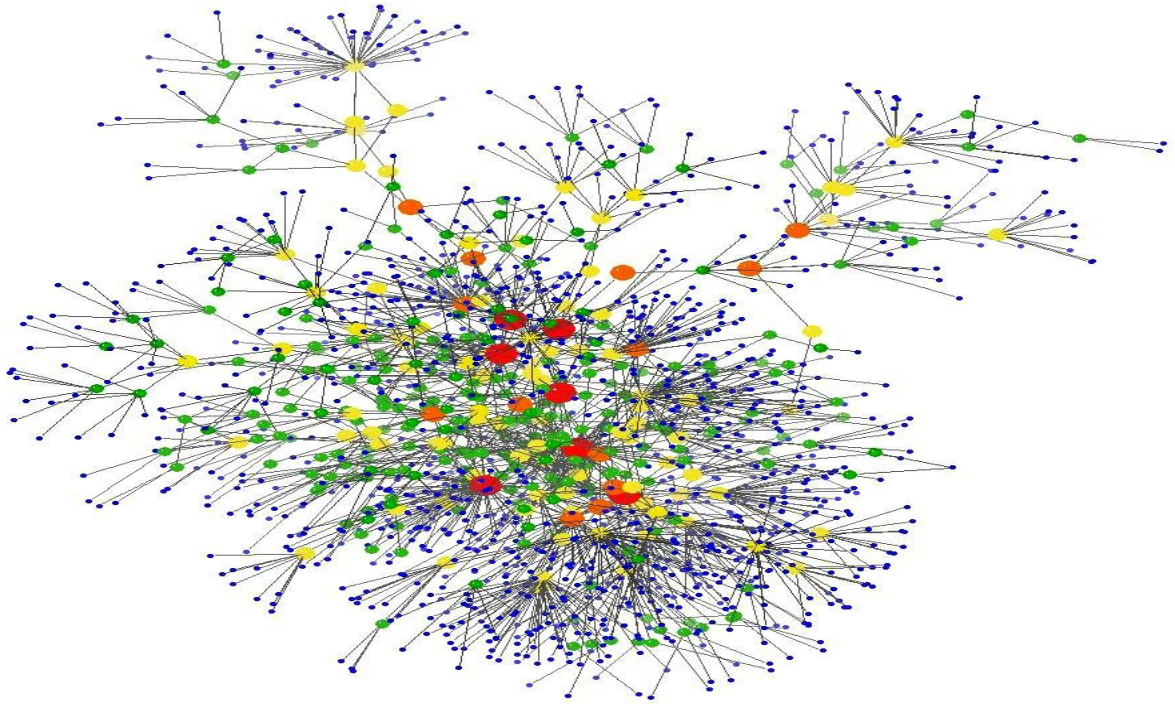


A model for data description : RDF model



A model for data description : RDF model

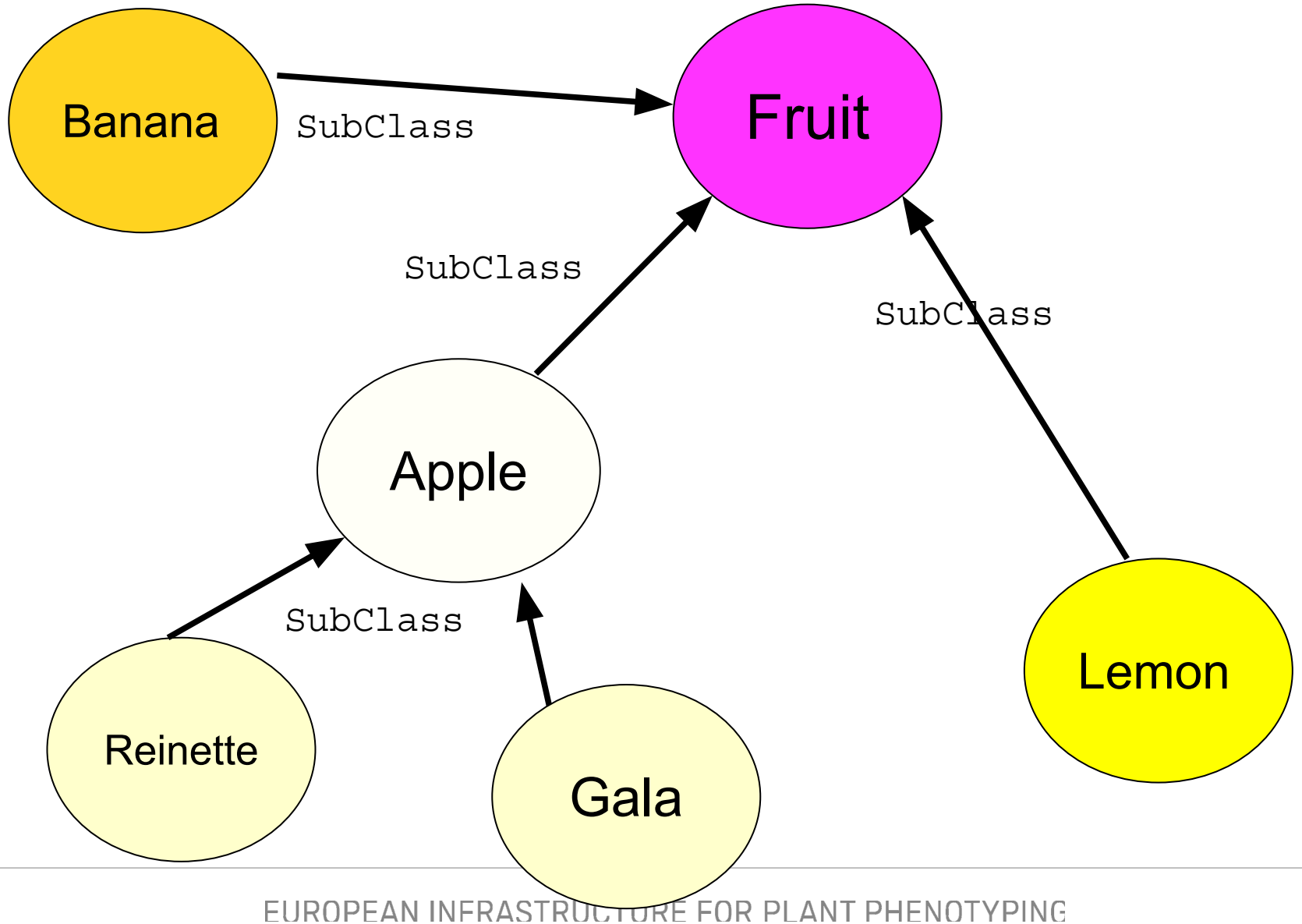
- Linked data
- Distributed semantic graphs
- Data organization & query



Resource structuring

- RDF Schema
  - Controlled vocabularies
  - Hierarchy of classes (class, sub-class, …)
  - Property domaine and range

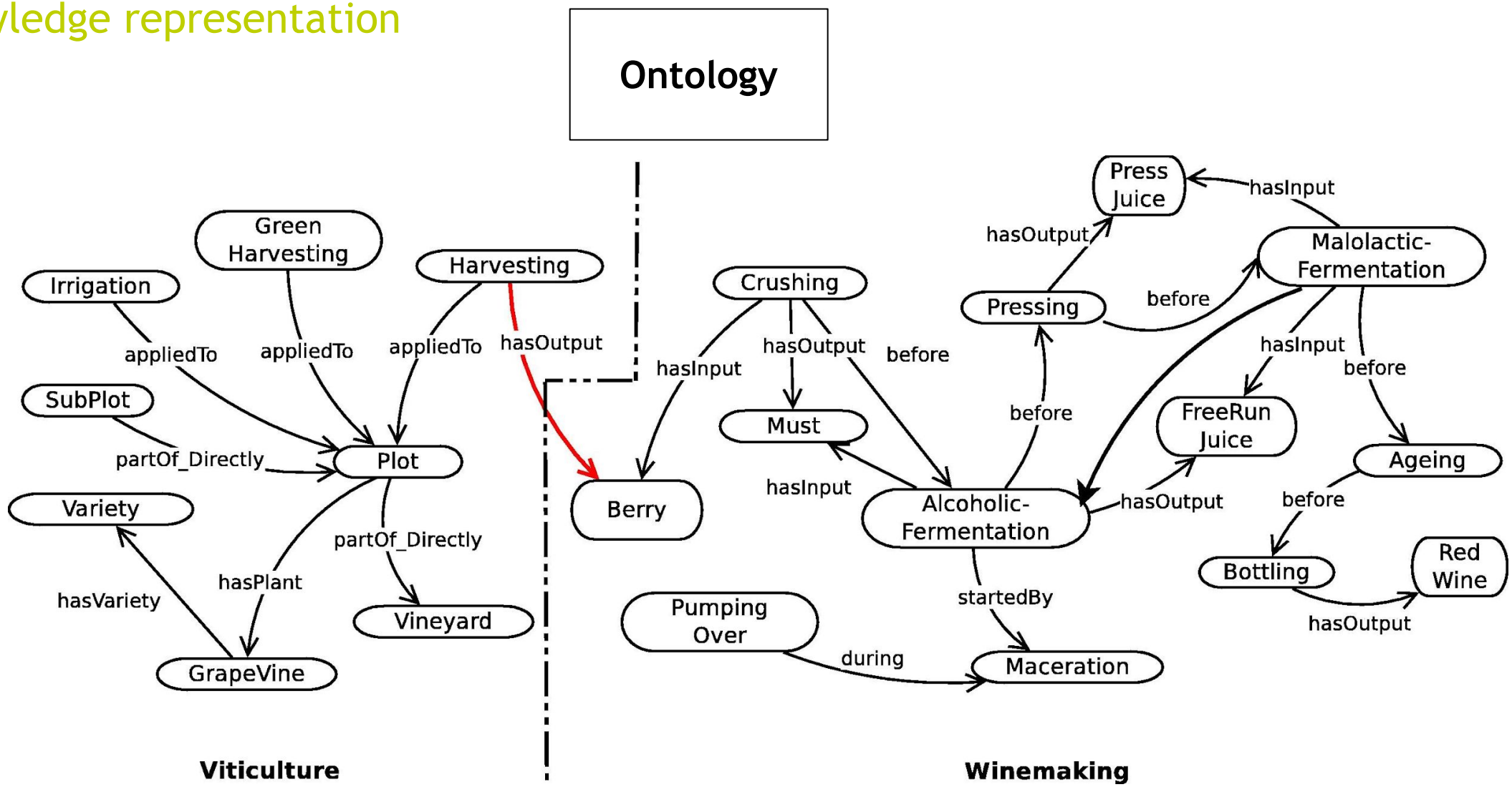
A model for data description : RDF model



Knowledge representation

- Ontology Web Language (OWL)
  - Knowledge representation about classes and/or properties
  - Based on logical descriptions and reasoning (rules & inference)
  - Integration of knowledge in information systems

Knowledge representation

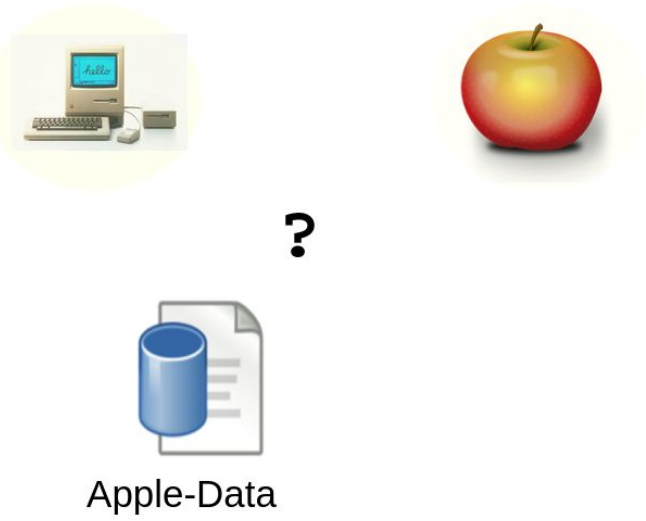


Semantic graph query

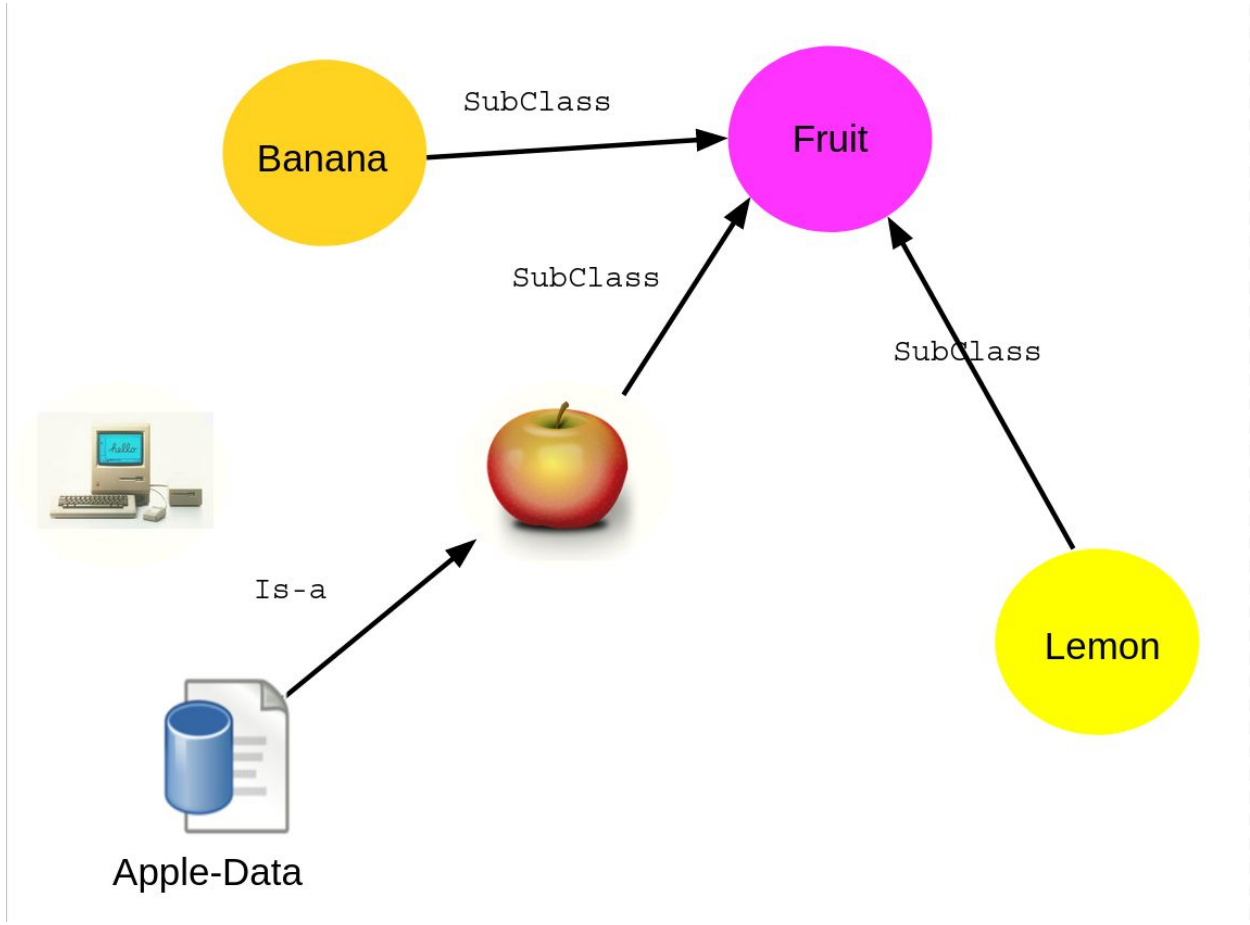
- SPARQL: A query language for RDF graphs
  - Search
  - Modify
  - Add
  - Delete

```
PREFIX foaf:
<http://xmlns.com/foaf/0.1/>
    SELECT ?name ?email
    WHERE {
    ?person a foaf:Person.
    ?person foaf:name ?name.
    ?person foaf:mbox ?email.
    }
```

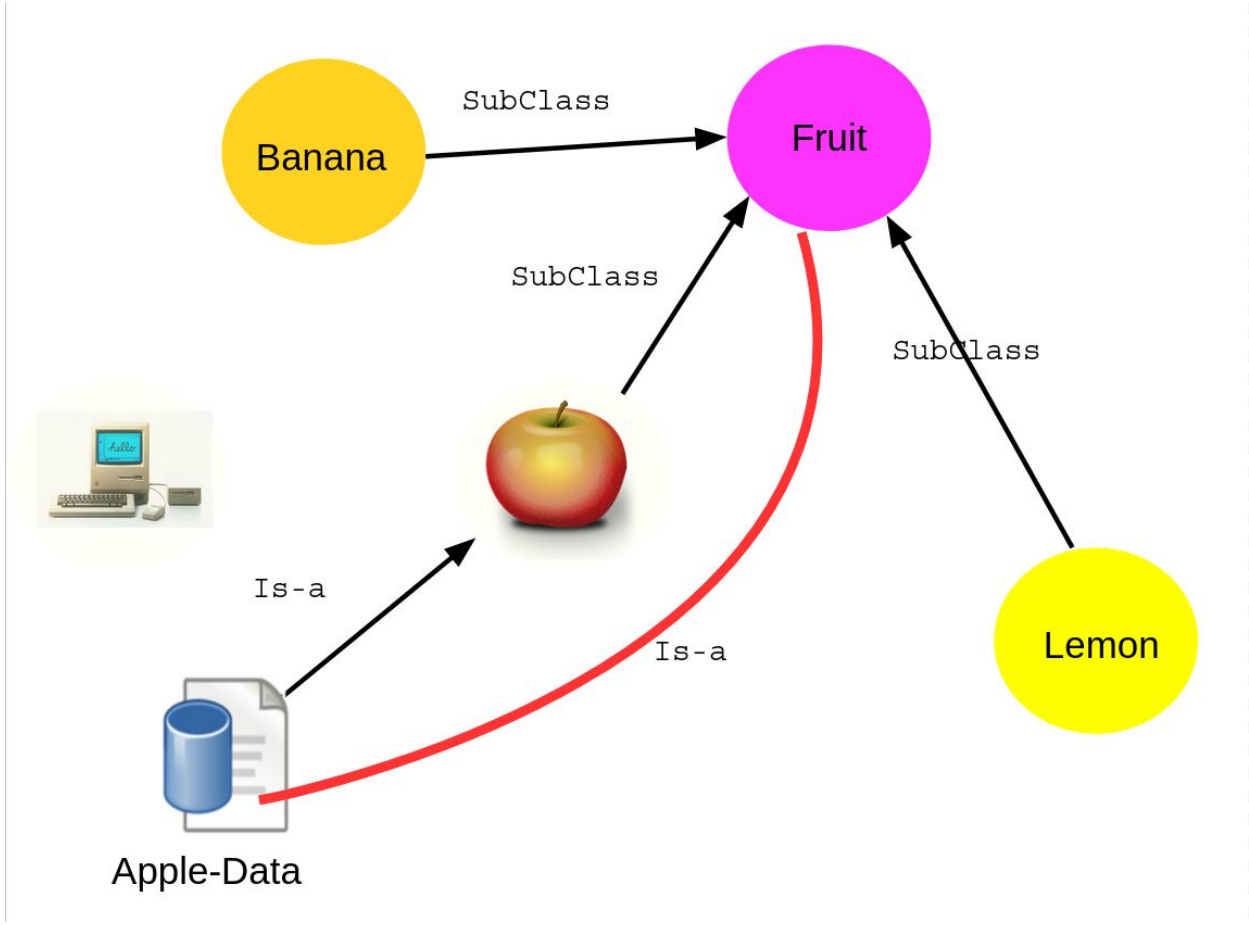
Semantic graph query



Semantic graph query



Semantic graph query

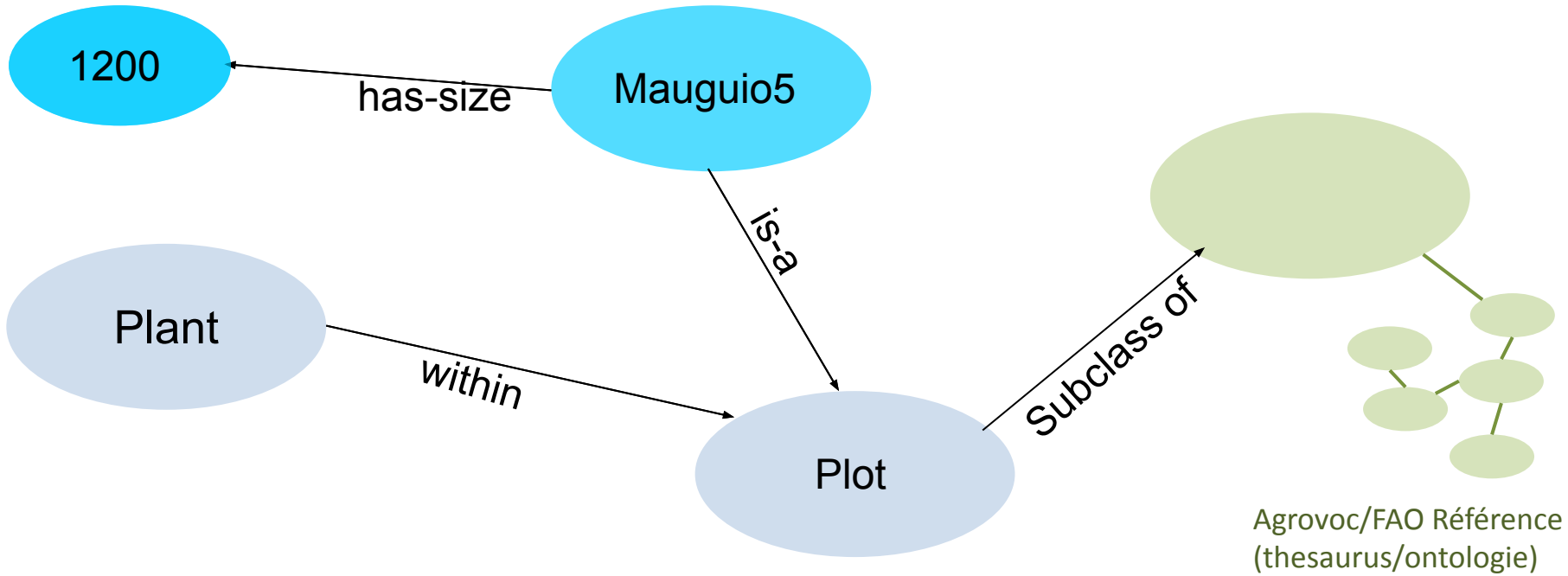


Controlled vocabulary

Idea:

- A controlled vocabulary is a predefined list of values to be used for specific properties
- Common controlled vocabularies to make data understandable across systems ⇒ machine readable
- Share & reuse via portals

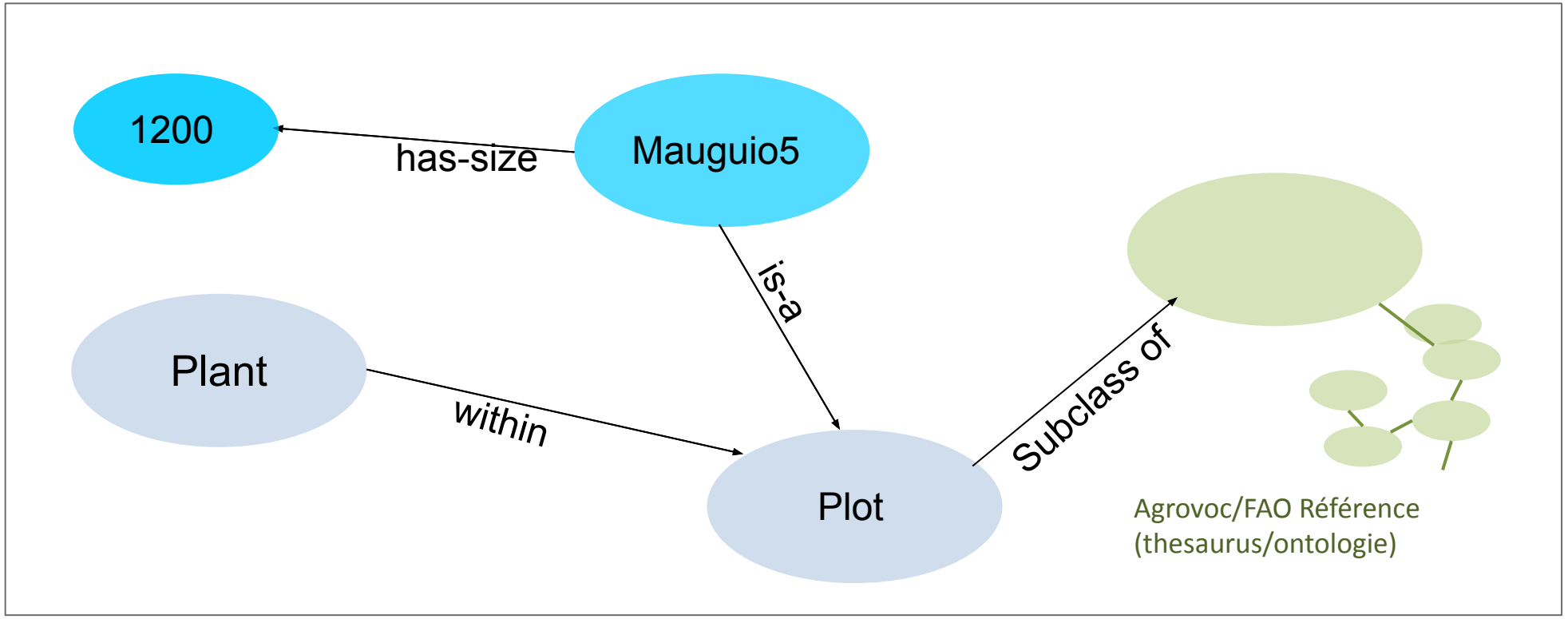
Semantic Web : Implementation and use of ontologies (OWL)



Semantic Web : Common framework to share and reuse data

⇒ Implementation and use of ontologies (OWL)

Ontology: Gives meaning to data ⇒ link each element of data to a controlled and shared vocabulary



Interoperability

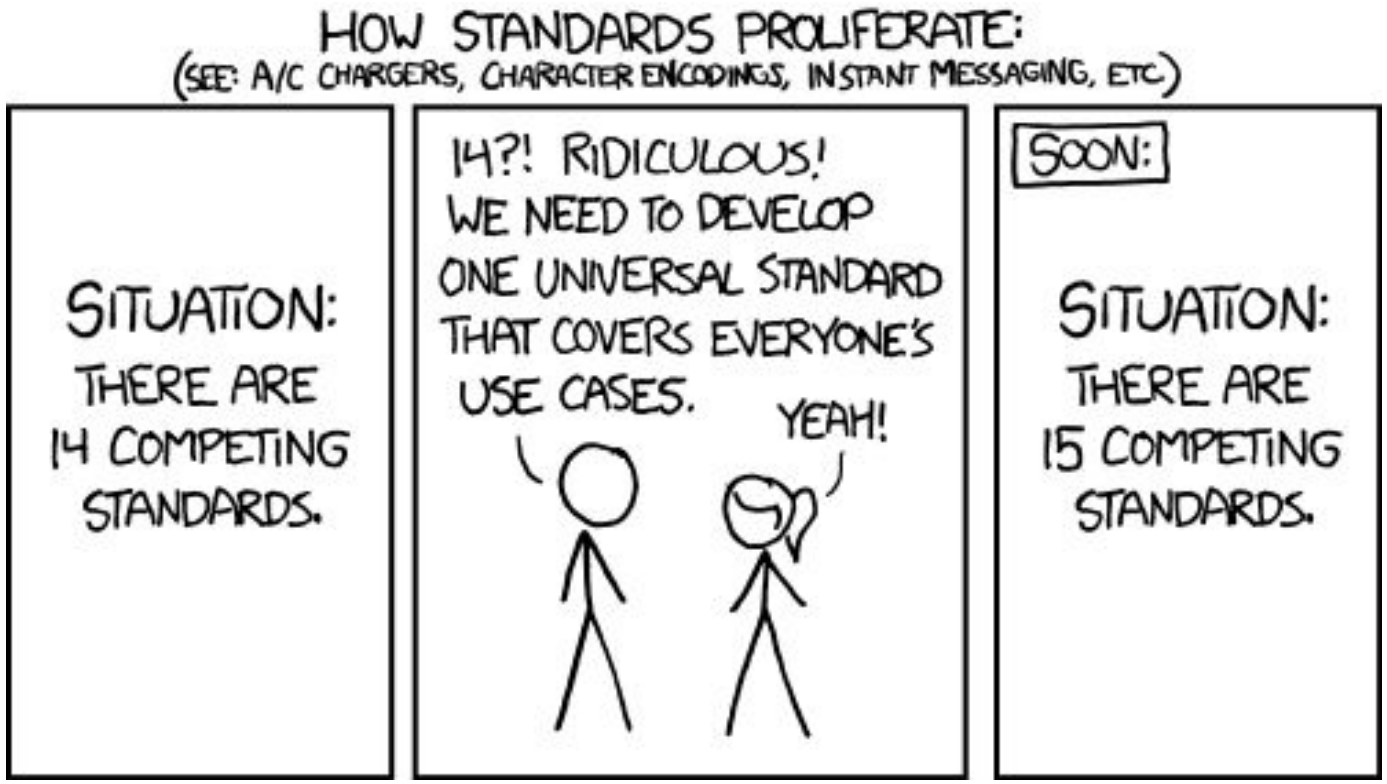
Ability of a computer system to work with another system (tools & standards)

- **Technologic (Communication abilities)**
- **Syntaxic (Communication skills)**
- **Semantic (Comprehension abilities)**

- ❖ Structure, content, integrity, …
- ❖ Make data searchable
- ❖ Perpetuate, reusable and shareable

Use standards

Standard adoption & perennity: Why using standards?



Interoperability in international network - Standards

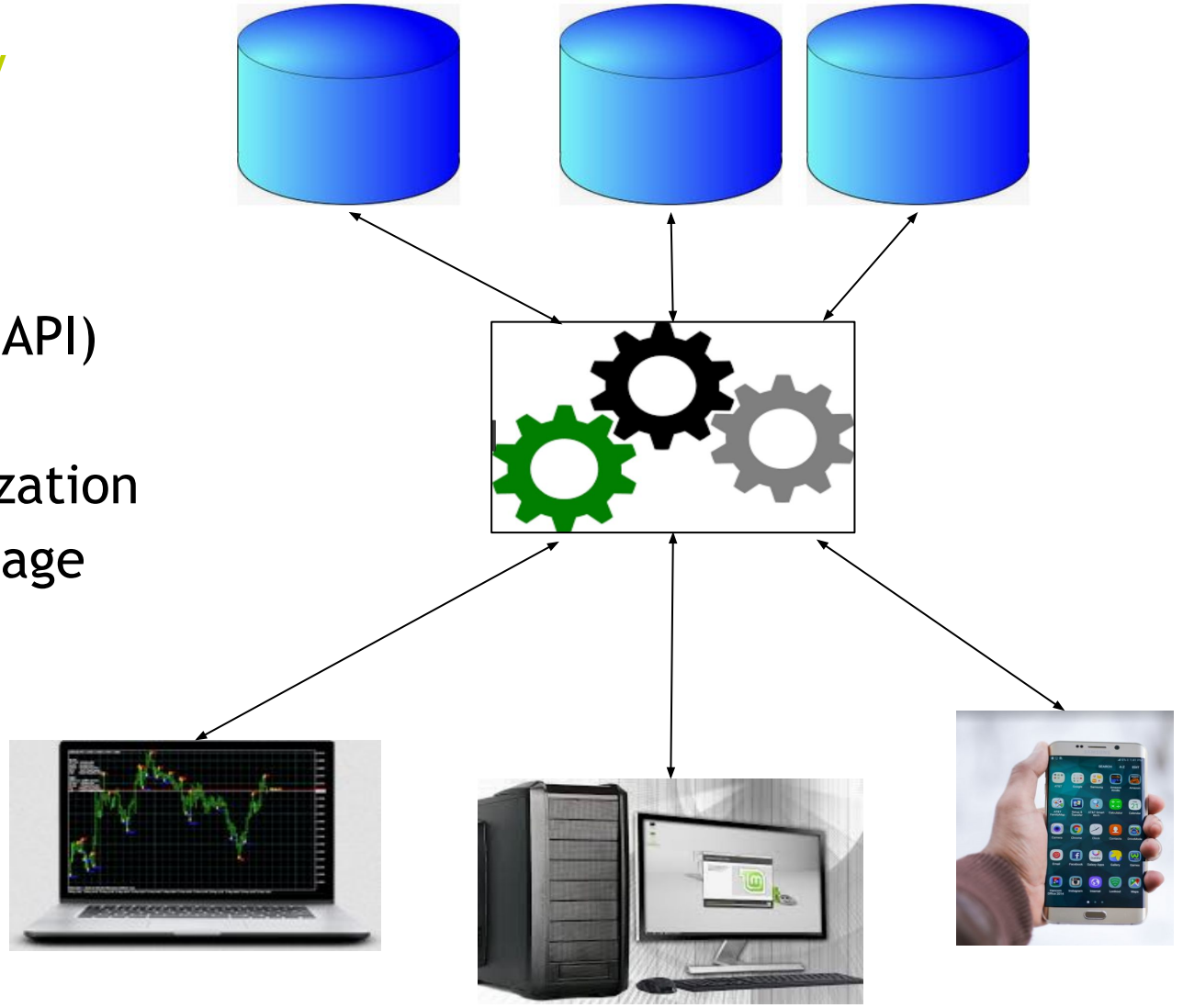
Some actors & Contributors

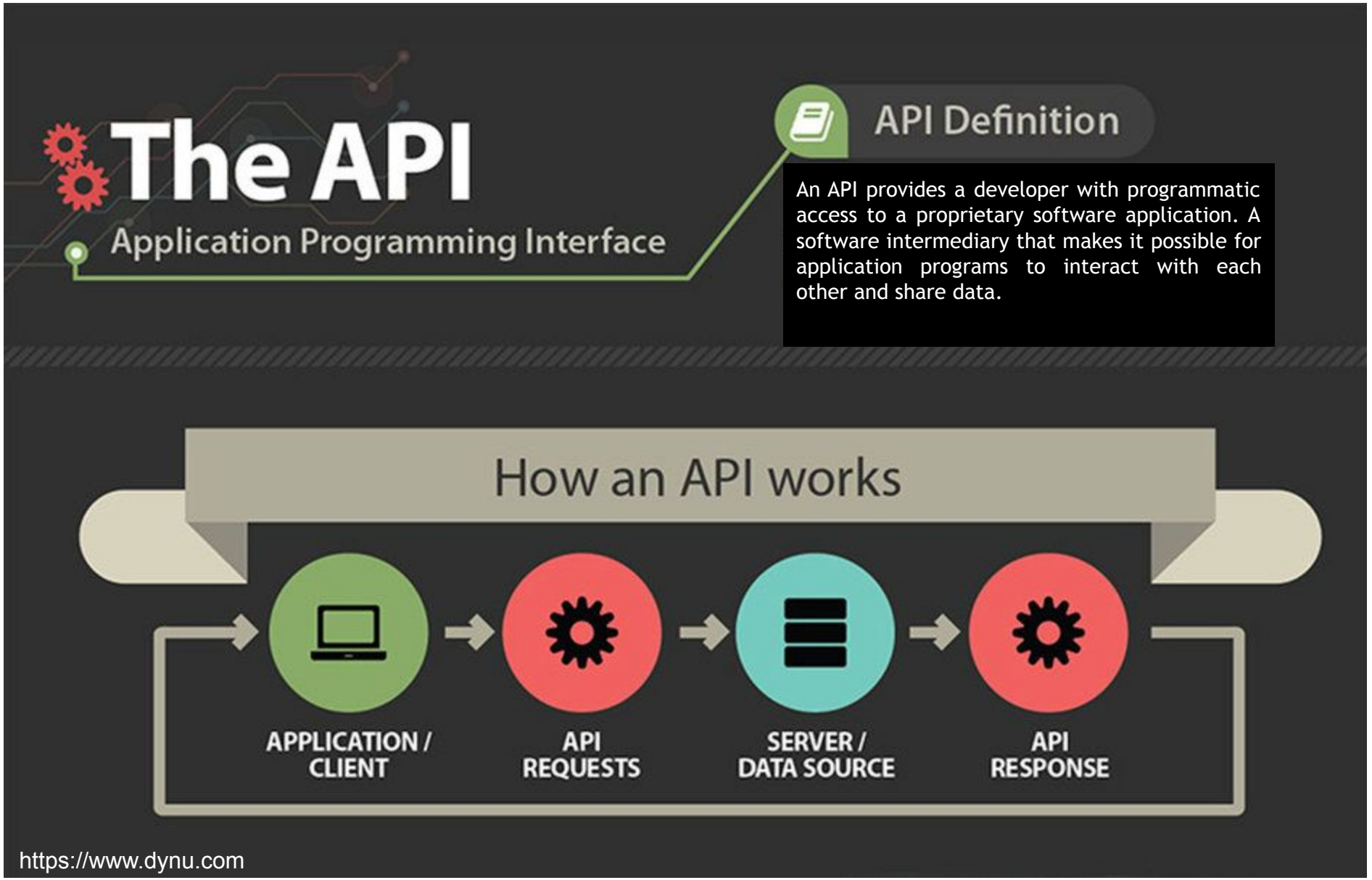
Some standards

MCPD

Web services: methods for data display

- Web for machines
  - Interface of data publication (Web API)
  - Software component
  - Abstraction of internal data organization
  - Independent of programming language
  - http (dealing with firewall)
  - Client-Server architecture

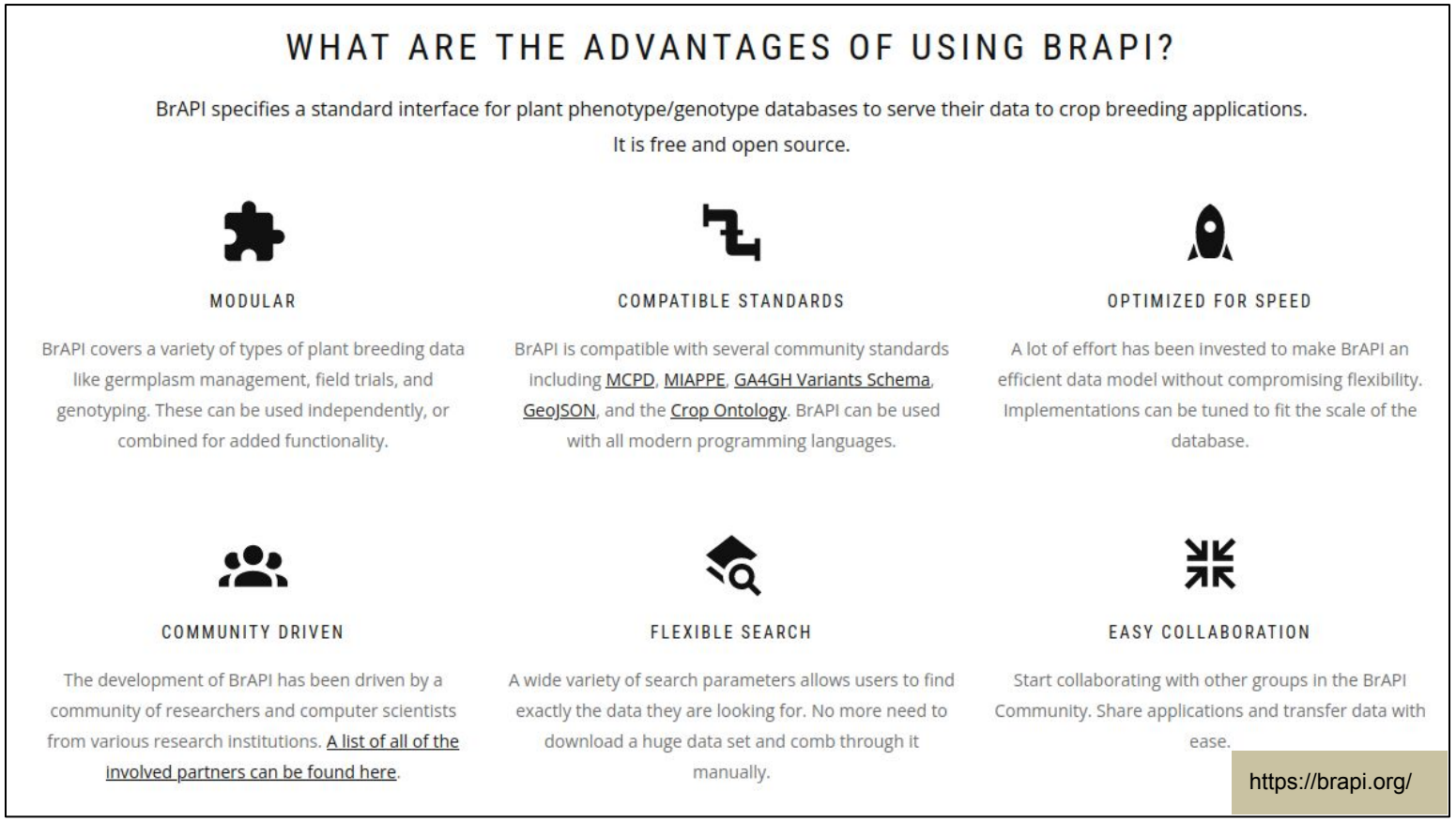



The API
Application Programming Interface
API Definition
An API provides a developer with programmatic access to a proprietary software application. A software intermediary that makes it possible for application programs to interact with each other and share data.
How an API works
APPLICATION / CLIENT
API REQUESTS
SERVER / DATA SOURCE
API RESPONSE
https://www.dynu.com

Examples of APIs used in plant sciences

BrAPI

Standardized RESTful Web Service API Specification for plant breeding data



WHAT ARE THE ADVANTAGES OF USING BRAPI?

BrAPI specifies a standard interface for plant phenotype/genotype databases to serve their data to crop breeding applications. It is free and open source.

MODULAR

BrAPI covers a variety of types of plant breeding data like germplasm management, field trials, and genotyping. These can be used independently, or combined for added functionality.

COMPATIBLE STANDARDS

BrAPI is compatible with several community standards including MCPD, MIAPPE, GA4GH Variants Schema, GeoJSON, and the Crop Ontology. BrAPI can be used with all modern programming languages.

OPTIMIZED FOR SPEED

A lot of effort has been invested to make BrAPI an efficient data model without compromising flexibility. Implementations can be tuned to fit the scale of the database.

COMMUNITY DRIVEN

The development of BrAPI has been driven by a community of researchers and computer scientists from various research institutions. A list of all of the involved partners can be found here.

FLEXIBLE SEARCH

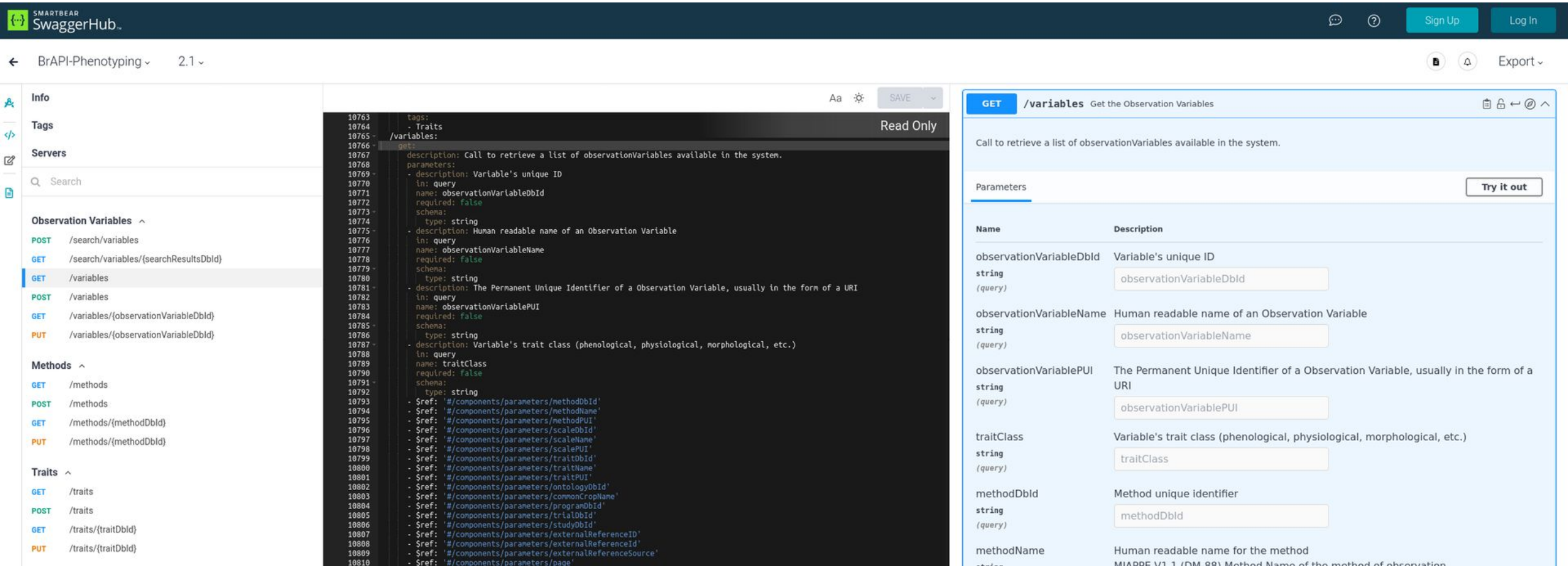
A wide variety of search parameters allows users to find exactly the data they are looking for. No more need to download a huge data set and comb through it manually.

EASY COLLABORATION

Start collaborating with other groups in the BrAPI Community. Share applications and transfer data with ease.

https://brapi.org/

BrAPI


SMARTBEAR SwaggerHub
Sign Up
Log In
BrAPI-Phenotyping
2.1
Export
Info
Tags
Servers
Search
Observation Variables
POST /search/variables
GET /search/variables/{searchResultsDbId}
GET /variables
POST /variables
GET /variables/{observationVariableDbId}
PUT /variables/{observationVariableDbId}
Methods
GET /methods
POST /methods
GET /methods/{methodDbId}
PUT /methods/{methodDbId}
Traits
GET /traits
POST /traits
GET /traits/{traitDbId}
PUT /traits/{traitDbId}
Aa
SAVE
Read Only
tags:
- Traits
/variables:
get:
description: Call to retrieve a list of observationVariables available in the system.
parameters:
- description: Variable's unique ID
in: query
name: observationVariableDbId
required: false
schema:
type: string
- description: Human readable name of an Observation Variable
in: query
name: observationVariableName
required: false
schema:
type: string
- description: The Permanent Unique Identifier of a Observation Variable, usually in the form of a URI
in: query
name: observationVariablePUI
required: false
schema:
type: string
- description: Variable's trait class (phenological, physiological, morphological, etc.)
in: query
name: traitClass
required: false
schema:
type: string
- $ref: '#/components/parameters/methodDbId'
- $ref: '#/components/parameters/methodName'
- $ref: '#/components/parameters/methodPUI'
- $ref: '#/components/parameters/scaleDbId'
- $ref: '#/components/parameters/scaleName'
- $ref: '#/components/parameters/scalePUI'
- $ref: '#/components/parameters/traitDbId'
- $ref: '#/components/parameters/traitName'
- $ref: '#/components/parameters/traitPUI'
- $ref: '#/components/parameters/ontologyDbId'
- $ref: '#/components/parameters/commonCropName'
- $ref: '#/components/parameters/programDbId'
- $ref: '#/components/parameters/trialDbId'
- $ref: '#/components/parameters/studyDbId'
- $ref: '#/components/parameters/externalReferenceID'
- $ref: '#/components/parameters/externalReferenceId'
- $ref: '#/components/parameters/externalReferenceSource'
- $ref: '#/components/parameters/page'
GET /variables Get the Observation Variables
Call to retrieve a list of observationVariables available in the system.
Parameters
Try it out
Name
Description
observationVariableDbId
string
(query)
Variable's unique ID
observationVariableDbId
observationVariableName
string
(query)
Human readable name of an Observation Variable
observationVariableName
observationVariablePUI
string
(query)
The Permanent Unique Identifier of a Observation Variable, usually in the form of a URI
observationVariablePUI
traitClass
string
(query)
Variable's trait class (phenological, physiological, morphological, etc.)
traitClass
methodDbId
string
(query)
Method unique identifier
methodDbId
methodName
Human readable name for the method

Conclusion : Data management in scientific research

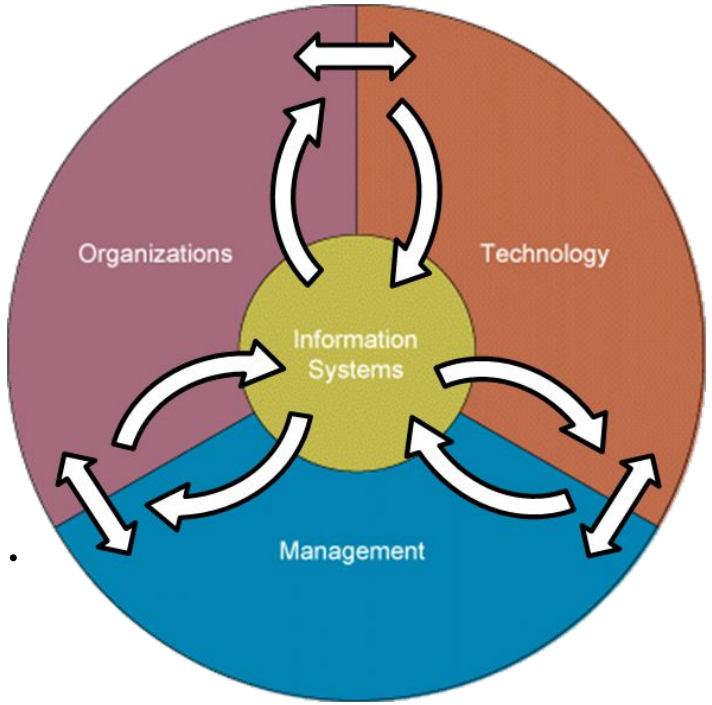
- We absolutely need Advanced Data Management :
  - Meta-Analysis, reuse, transparency
  - Interdisciplinary & participatory research

  However:

  - Still quite reluctant to share data
    - How data will be used and analysed?
    - Who will use these data?
  - Sometimes not able to reuse own team data (highly customized data sets)
  - Very busy collecting and analysing, no time for data management

What do we need?

- Data base management system (SGBD, NoSQL, RDF storage, etc)
- Representation language XML or JSON
- Use standards
- Access (Web pages, Web Services)
- Knowledge representation : ontology, thesaurus, taxonomy, …
- Analysis and visualisation tools (R, python, ….)



Thank you for your attention!

https://www.phenome-emphasis.fr/

OpenSILEX Team - http://opensilex.org/

Special thanks to: Silvana Moscatelli, François Tardieu, Cyril Pommier, …

EMPHASIS

emphasis@fz-juelich.de

emphasis.plant-phenotyping.eu

EMPHASIS_EU

EMPHASIS.EU

EMPHASIS on Plant Phenomics

ESFRI

EMPHASIS is an ESFRI-listed project.

EMPHASIS-PREP is funded by the
European Union (Grant Agreement: 739514).

EUROPEAN INFRASTRUCTURE
FOR PLANT PHENOTYPING