Building a consistent Information System in the different nodes and defining standardisation strategies

Romain David, Jean-Eudes Hollebecq, Pascal Neveu

European
Plant Phenotyping
Network
EPPN 2020

2nd Annual Meeting May 2019

WUR - NL

Romain David - Jean-Eudes Hollebecq - Pascal Neveu, INRA

JRA3

‘Building a consistent Information System in the different nodes and defining standardisation strategies’

- **Object identification**
  Objects: plants, plots, experiments, sensors, events, etc
  Identification: persistent, unambiguous, resolvable

- **Variable naming and formalization**
  Give (local or global) name to variables
  What (concept), How, Associated controlled contexts

- **Data interoperability**
  Formats, schemas, semantic
  Representation compatibility and consistency

Object Identification

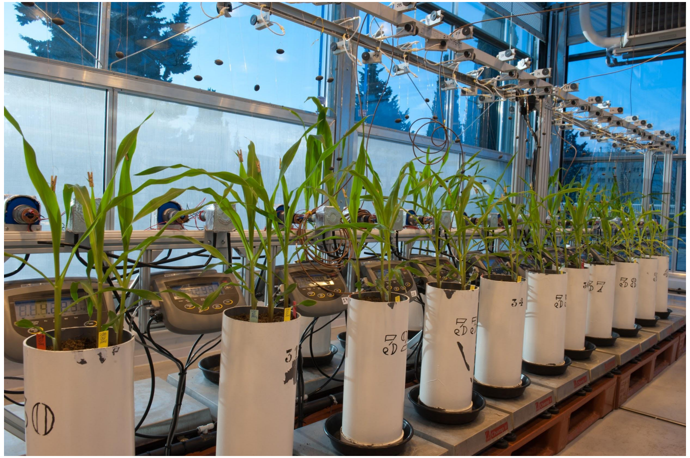
What do we want to identify?

Organisational objects and actors:
Organisations, Experiments, Locations, Operators, Softwares, etc.

Biological material:
Plants, Leaves, Stems, Flowers, Roots, etc.

Omics data:
Proteins, Spectrum, Transcriptum, etc.

Experimental Material:
Sensors, Pots, Substrats, etc.

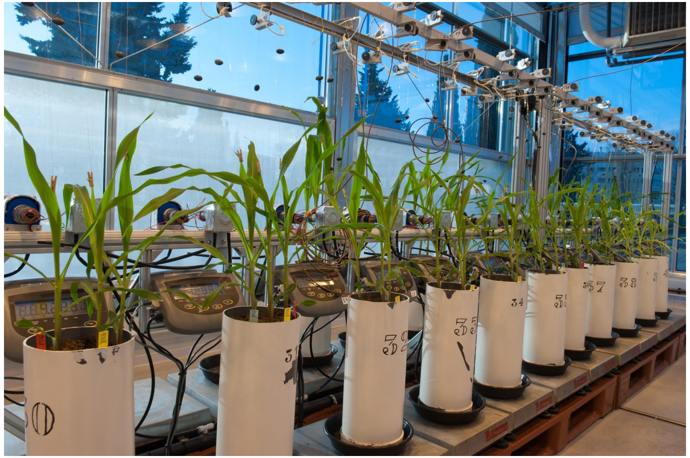


Object Identification

What do we want to identify?

Collaborative objects, resources:
Aggregated data, Concepts, Sample-analysis, etc

Events and activities
Faults, Management, Disturbance, Meteo, etc.



Digital resources :
Datasets, Reports, Papers, etc.

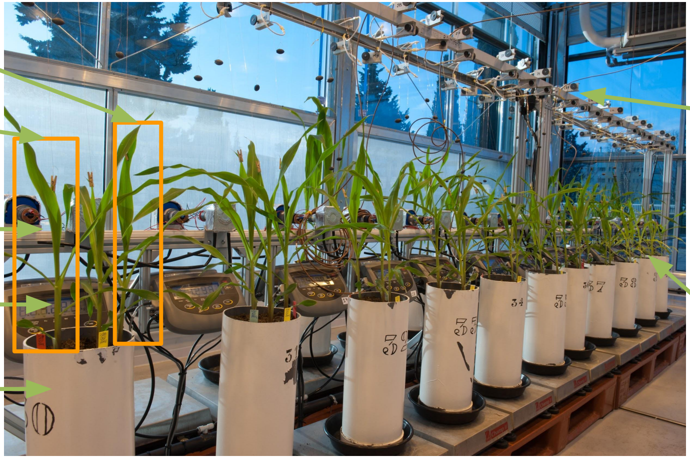
Object Identification

Concepts / **Classes**: Pot, Plant, Leaf

Instances: pot22, pot17, leaf234, plant12

INRA-LEPSE

greenhouse22



plant13

plant12

leaf234

stem234

pot22

sensor628

soilAB

Object Identification

What is an identifier?

An identifier is a sort of name (could be alphanumeric or numeric only) that identifies a specific object (digital or not) in a set of objects.

DOI or URI are string that identifies a particular resources

DOI: 10.1111/nph.15385 → http://doi.org/10.1111/nph.15385

```
http://www.inra.fr/mp3/2015/arch/exp21/plant227
```

In an ideal world identifier should be **unique for each object (bijection)**,

In practice this is rarely the case.

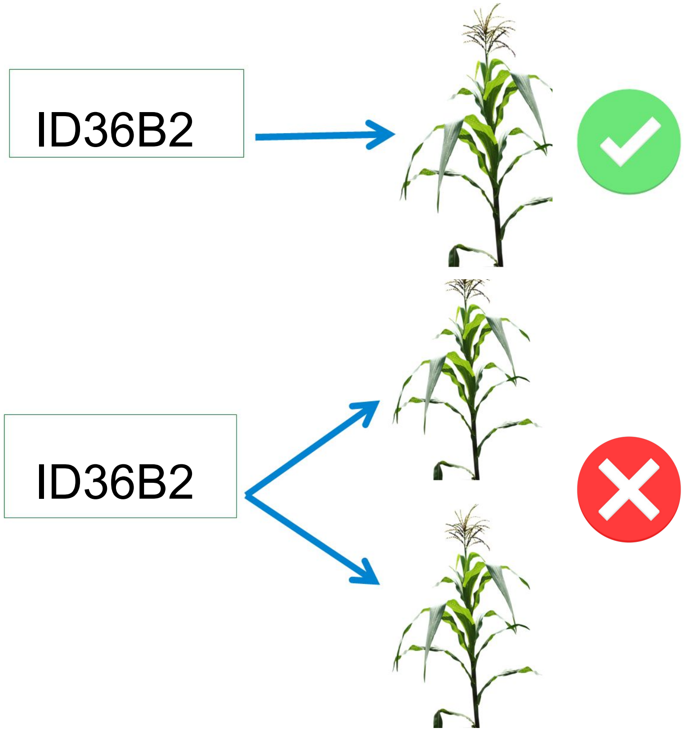
Object Identification

How do we want to identify? good **identifier**.

Non ambiguous
An identifier only stands for one resource. Whatever the database or source, two objects can not have the same identifier.

Confusions happens when different resources share the same identifier.

This characteristic is mandatory for identifiers.



Object Identification

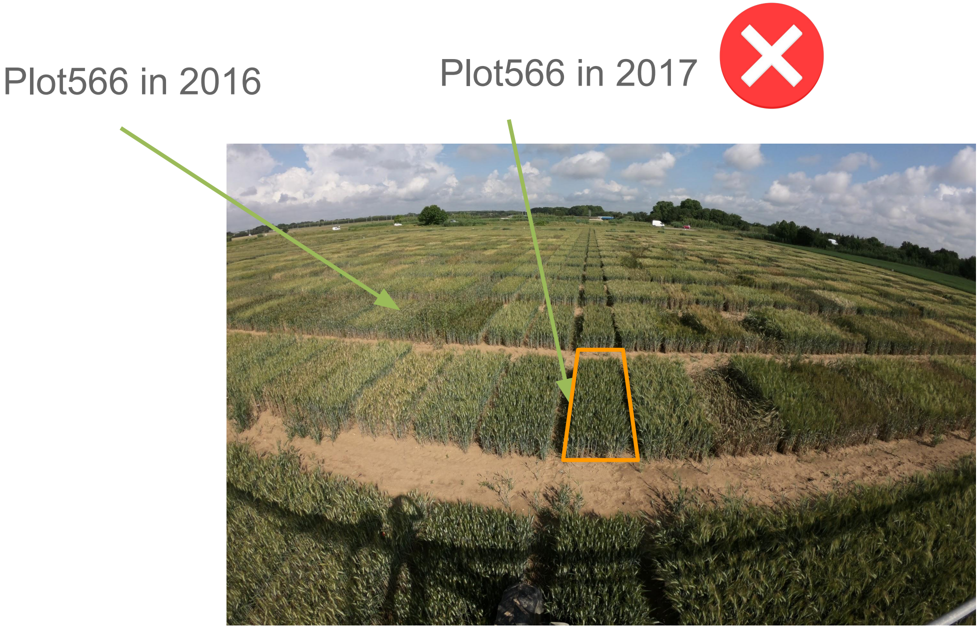
How do we want to identify? good **identifier**.

Non ambiguous

An identifier only stands for one resource. Whatever the database or source, two objects can not have the same identifier.

Confusions happens when different resources share the same identifier.

This characteristic is mandatory for identifiers.



can change over time:
e.g. plot cutting

Object Identification

How do we want to identify? good **identifier**.

Non ambiguous
Persistent

A persistent identifier is an identifier that is **permanently assigned to an object** (Ideally usable in several decades).

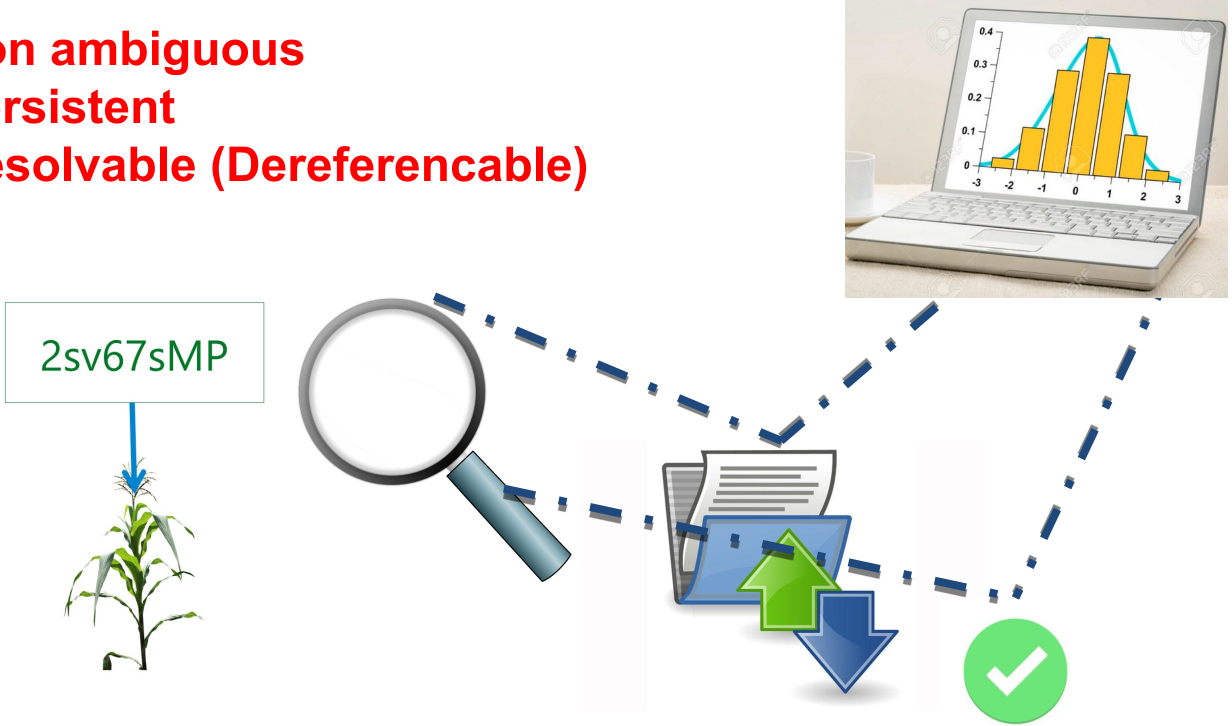
Aims : reusability of data over the long term (H2020 requirement)

The problem is that during periods of decades, many changes can occur within databases, but also in institutions or organizations in charge of the data.
It is thus necessary to preserve and recover dependencies between these elements, this in time and localisation.

Object Identification

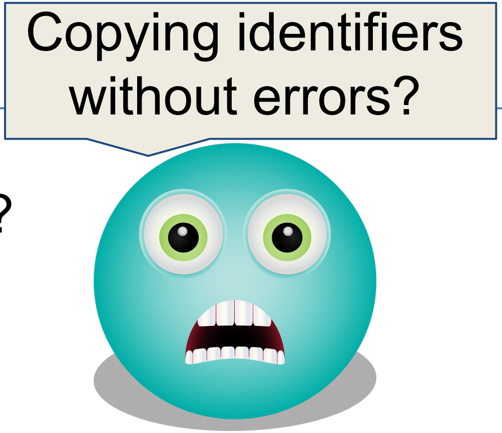
How do we want to identify? With **an efficient identifier**.

Non ambiguous
Persistent
Resolvable (Dereferencable)



An identifier is said to be dereferenceable if it is possible to access the object or all the digital contents describing the object (e.g. URL, URI...).

Identification

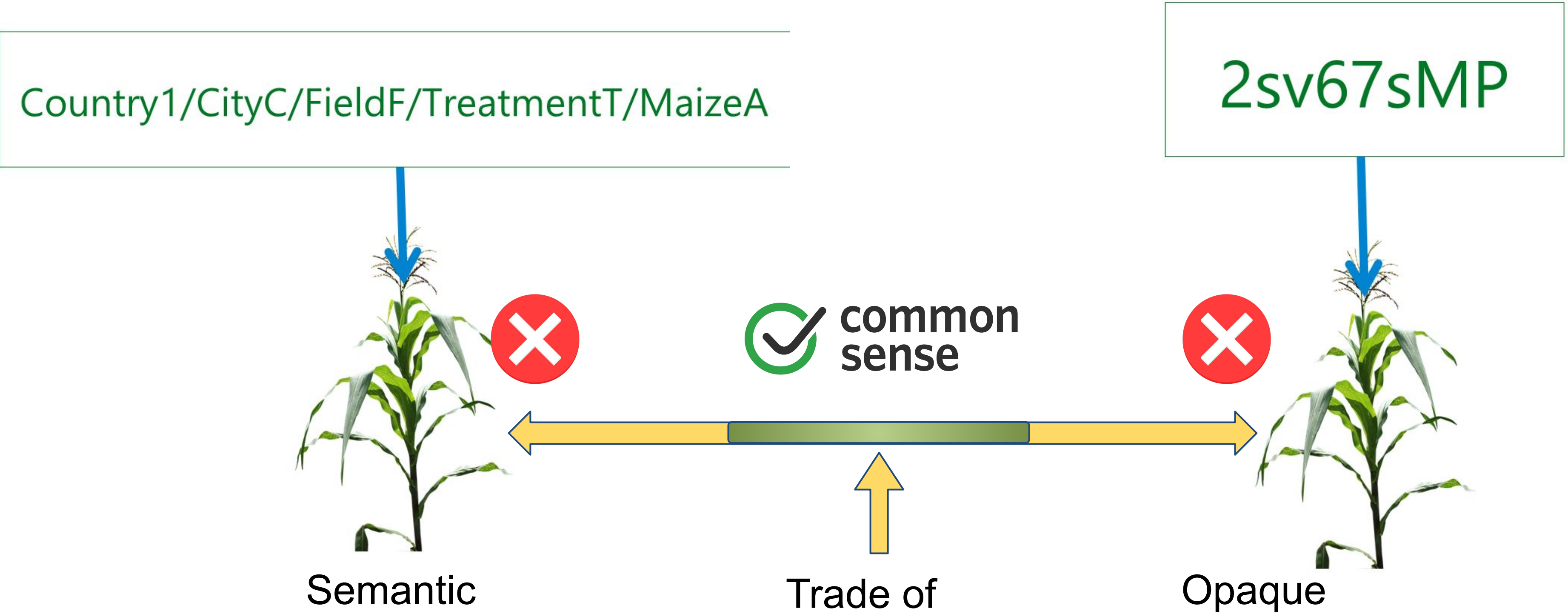


What do <u>we want **to avoid**</u> for an efficient identification?

Non usability for operators:

- No semantic in the identifier / excessively long
  - 10908538265365831680853826536583168
  - http://www.domainname.fr/m3p/arch2017c17000915/FMfcgxwCgCSgPwdZxzkZtSHhnmvWWcpp
- Confusing letters
  - Little L and maj i : l & I are not easily differentiable
  - (*idem* for O letter and 0 number)
- Too much semantic in the identifier (avoid optional metadata)
  - (http://www.domainname.fr/m3p/Program/work_package/Country/Site/Year/Month/Day/Operator/Plant/Leave/Method/Color_of_the_pen/Length)

Identification



Country1/CityC/FieldF/TreatmentT/MaizeA

2sv67sMP

common sense

Semantic

Trade of

Opaque

Identification

Summary of what we want for an efficient identification

- **Non ambiguous**
- **Persistent (based on data authority)**
- **Potentially resolvable** (Dereferencable / web compatible)

Agility needed for the long term maintenance!!

And

- **Only one language** (english, other languages can be generated by alias)
- **Minimum recognizable semantic part**
  (Human readable to know unambiguously what it is during manipulations)
- **Not too long**
- **Easy to generate and use**

Data authority: a community approved and identified institution or body that is responsible of any type of action concerning data on the long term.

Variable naming

Variable naming and variable representation

- Measuments / observations / aggregated / calculated

potentially associated to all object types: biological material, events, organizational objects

And be

- unidimensional or multidimensional
- quantitative, qualitative, symbolic (intervals)
- spatial, temporal, thermal time
- phenotypic, environmental

Naming rules

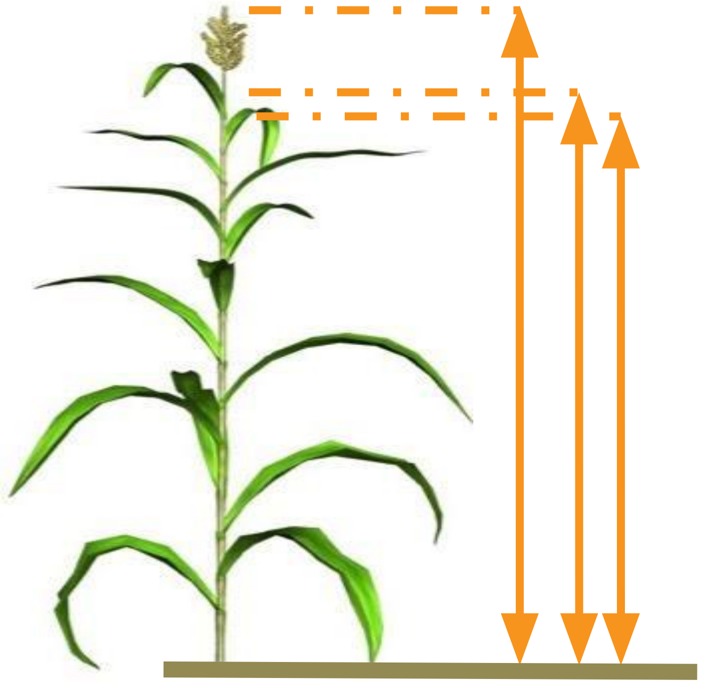


Figure From D. Pot, CIRAD

Plant Height

How to define a variable?

The example of *Plant Height*

- Plant with root?
- Plant with flower?
- Stem and leaves?
- Only the stem?
- Dry or not?
- In the morning or the end of the day?

Variable naming

What do <u>we want **to avoid**</u> in global context

- several names for same variable
- same name for several variable ⟶
- Sharing fuzzy or unstable variables
- Zero semantic name
- Beware I (i) l (L) O and 0 (zero) in names
- No optional metadata in names

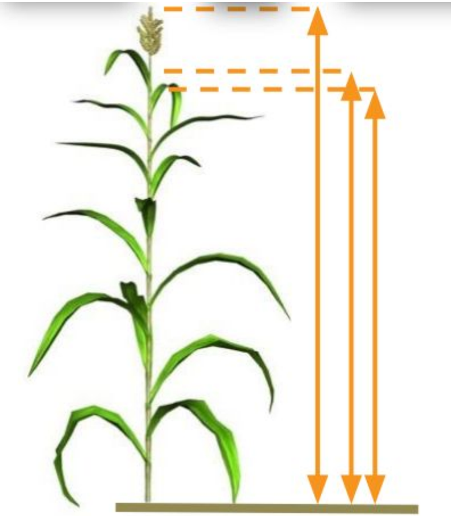


Figure From D. Pot, CIRAD

Variable naming

What do we recommended in global context

- Unambiguous name (in global context)
- Accessible description of variable
  Description can be read by machine and human
- Try to reuse existing variable if available
- Use standardized/shared representation schema for formalisation of new variable (and share it)

Variable naming

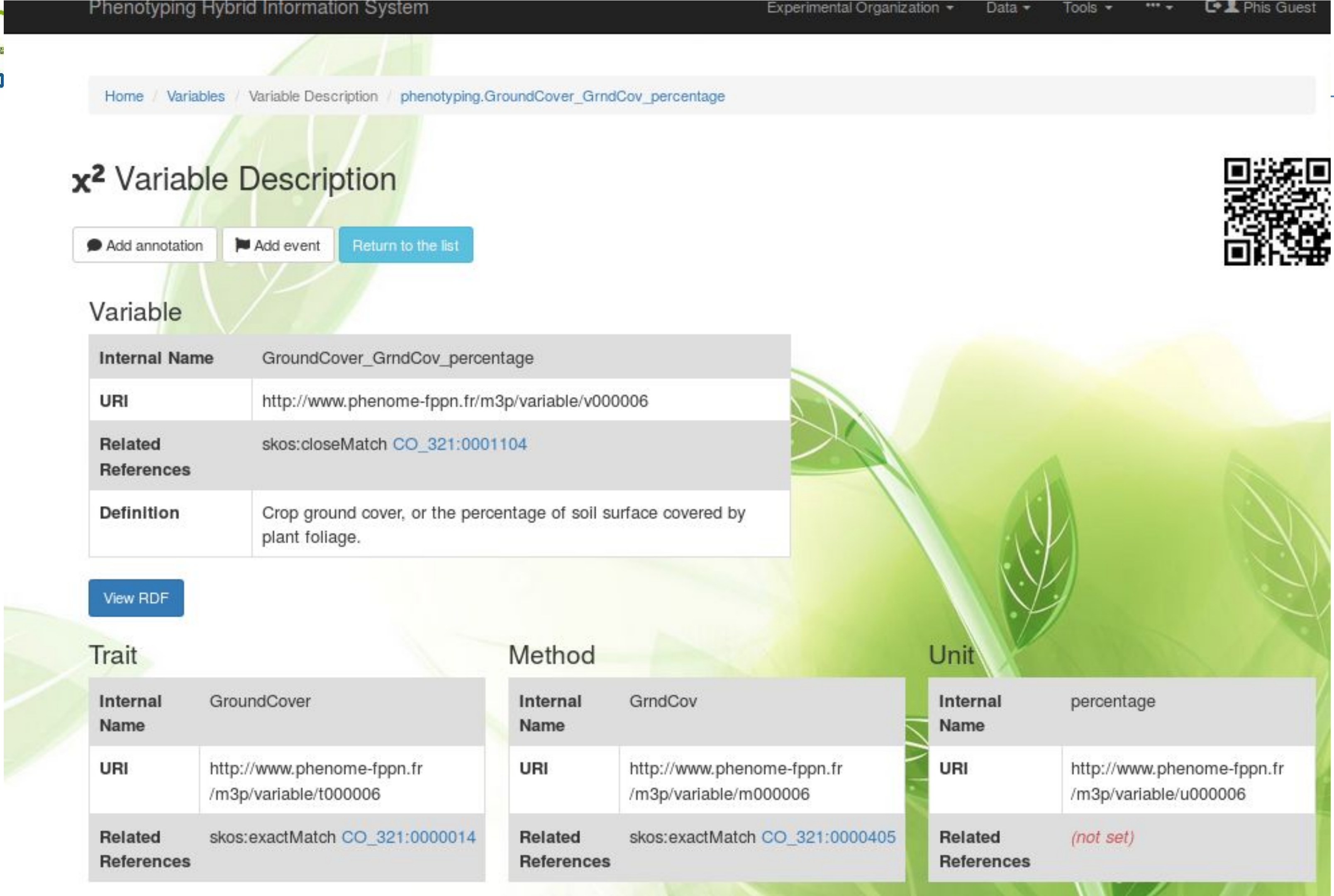
Variable representation schema for phenotypic variable

Variable = Trait + Method + Unit

Trait = Entity + Attribute (or Quality)

Entity, Attribute, Method and Unit must be referenced (if possible) in references ontologies and semantic resources:

Crop O., Trait O., Plant O., PATO, Agrovoc, ENVO, etc



Phenotyping Hybrid Information System
Experimental Organization
Data
Tools
Phis Guest

Home / Variables / Variable Description / phenotyping.GroundCover_GrndCov_percentage

x² Variable Description

Variable

Internal Name	GroundCover_GrndCov_percentage
URI	http://www.phenome-fppn.fr/m3p/variable/v000006
Related References	skos:closeMatch CO_321:0001104
Definition	Crop ground cover, or the percentage of soil surface covered by plant foliage.

Trait

Internal Name	GroundCover
URI	http://www.phenome-fppn.fr/m3p/variable/t000006
Related References	skos:exactMatch CO_321:0000014

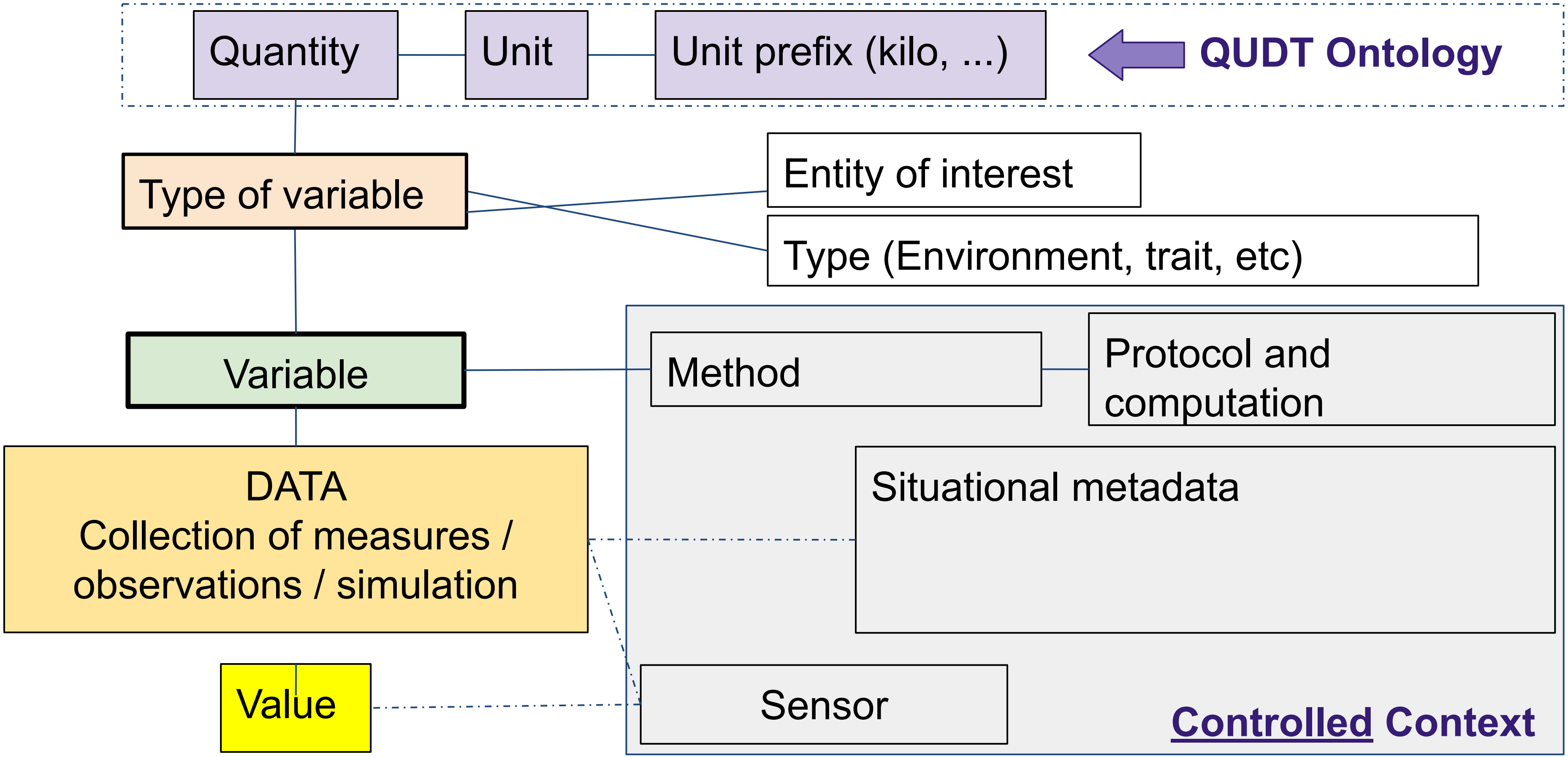
Method

Internal Name	GrndCov
URI	http://www.phenome-fppn.fr/m3p/variable/m000006
Related References	skos:exactMatch CO_321:0000405

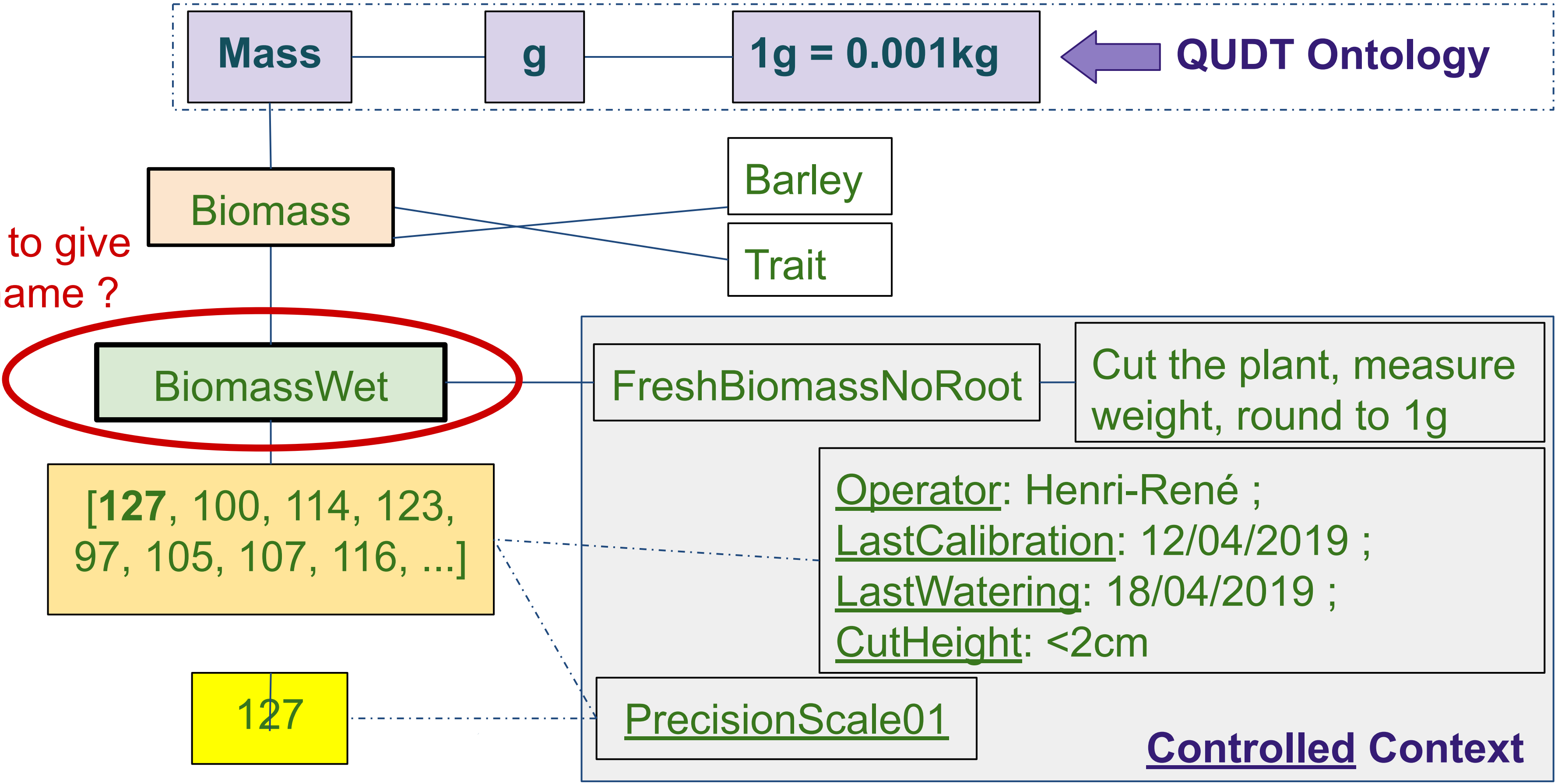
Unit

Internal Name	percentage
URI	http://www.phenome-fppn.fr/m3p/variable/u000006
Related References	*(not set)*

Naming rules - concepts



Naming rules



How to give it a name ?

Cut the plant, measure weight, round to 1g

Operator: Henri-René ;
LastCalibration: 12/04/2019 ;
LastWatering: 18/04/2019 ;
CutHeight: <2cm

Naming rules

What we could use

- Naming based on URI
    - share representation schema
    - Poor semantic URI
- Shared representation schema
- for “What”, “How”, “Context”, “Dimension”, etc
  (e.g. “height” of a “plant”)
- A part of the method / context / other
- can appear in aliases if necessary
    - Plant_Height_Met12
    - Plant_Height_Stem
      Plant_Height_StemRoot
    - ...

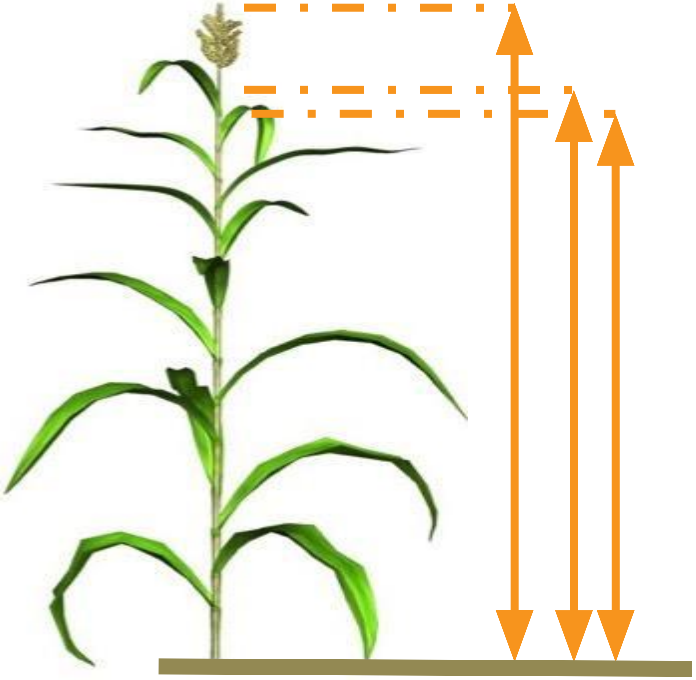


Figure From D. Pot, CIRAD

Plant Height

Naming rules

What do we want to control in the name (or in alias)?

Importance of a controlled context

Example of an UAV for remote sensing in the field

Controlled context data

- X, Y, Z, time, speed
- Sensor type (RGB, NDVI, etc.)
- Wind mean measurement

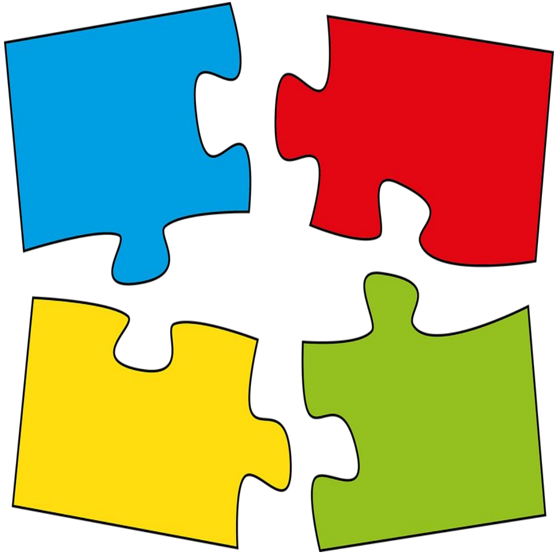
Not controlled context: Cloud, variability of wind

Not controlled context **affect the measure** (example of the clouds for the light)

Interoperability

Interoperable Informations systems allow data exchange and reuse among scientific disciplines, organisations and countries through **syntactically** parseable and **semantically** understandable operations.

.



Interoperability

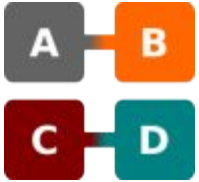
Interoperability ?

Improvement needs:

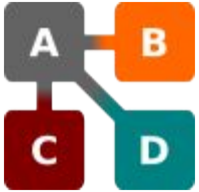
Data format
Data typology
Shared semantics
Rate of database updates

- Architecture
- Data qualification
- Taxonomic framework
- Repository of actors
- Conditions of use
- Accessibility
- Geographical repository
- …

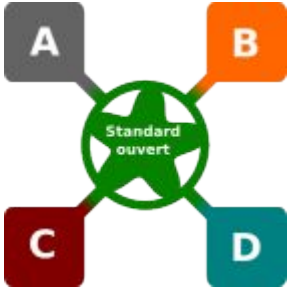
Compatibility



Standard

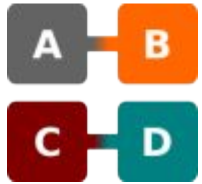


Interoperability



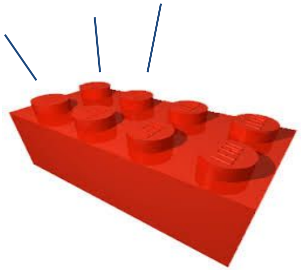
Interoperability

Bad practices ?

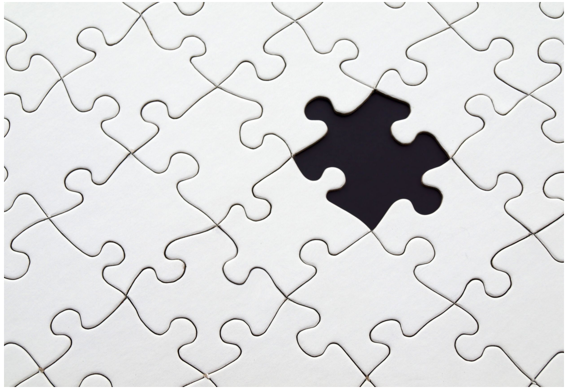


A & B are compatibles in another way than C & D :

Harder now to interoperate A & B with C & D



?



Interoperability is not only with your domain:

Do not create your own format / ontology without prospecting communities approved standards (in & outside of your community)

Interoperability

Improvement locks:

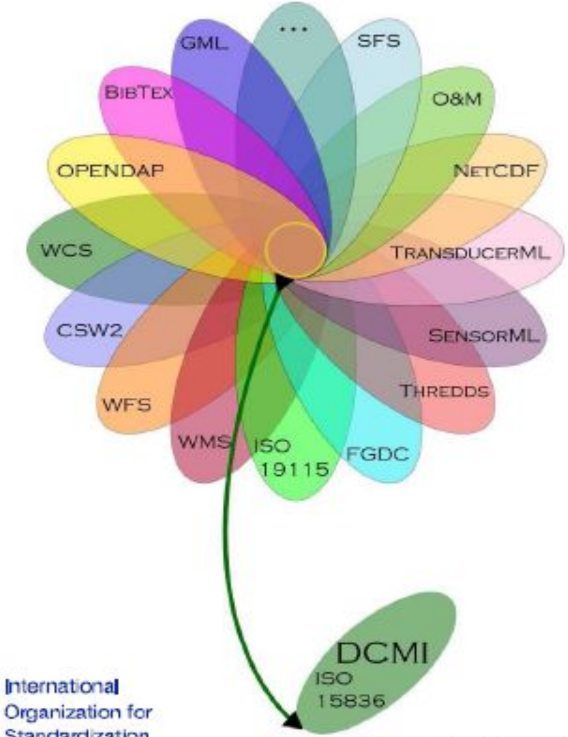
Too much standards and speed of standard evolution

Data format, Data typology, semantics & ontologies,
Data qualification, Data standards…
tools, technologies, groups...

OGC Open Geospatial Consortium, Inc.

W3C

ISO International Organization for Standardization



Source : Julien Barde IRD

$

Data cost

!

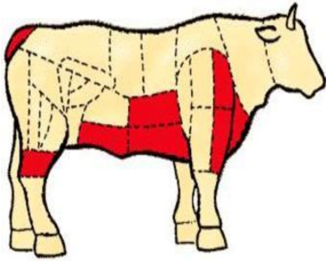
Skills availability

Interoperability

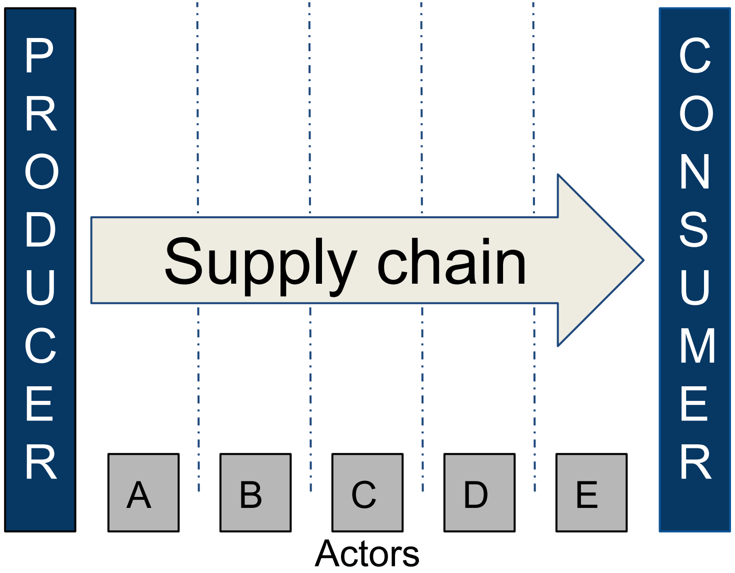
Interoperating in a safe way:

Working on interoperability also needs to work on traceability

Provenance → **Data tra**ceability is the capability to trace data and **each stage of their transformation**.



Challenge :
Identifying
-100% objects,
-100% actors,
-100% transformations



Interoperability

Provenance (data traceability) challenge

Provenance is the capability to trace data and **each of their transformation**.

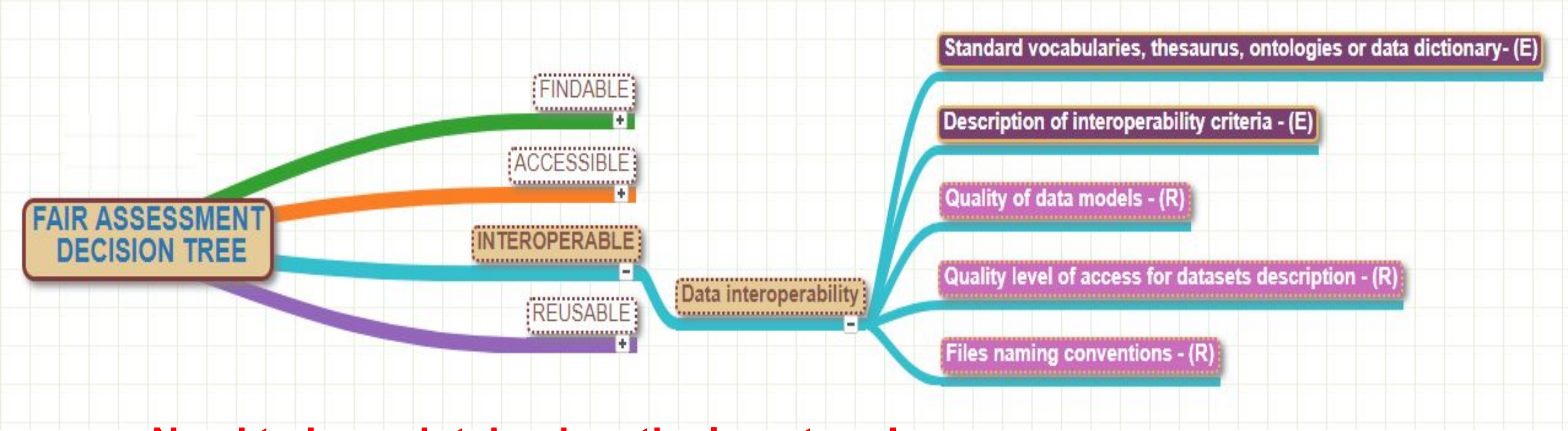
- ability to verify the
  - History / origin
  - location
  - curation
- keep track of a given set of information
  - in several distant information systems
  - within time
- keep track of the different versions

- by means of
  - Well documented
  - Metadata availability
  - Unambiguous identification
  - Naming

Interoperability

Interoperating in an efficient way:

Working on quality level of interoperability


FAIR ASSESSMENT DECISION TREE
FINDABLE
ACCESSIBLE
INTEROPERABLE
REUSABLE
Data interoperability
Standard vocabularies, thesaurus, ontologies or data dictionary- (E)
Description of interoperability criteria - (E)
Quality of data models - (R)
Quality level of access for datasets description - (R)
Files naming conventions - (R)

Need to be maintained on the long term!

take into account the persistence of human resources and skills

Romain David, Laurence Mabile, Mohamed Yahia, Anne Cambon-Thomsen, Anne-Sophie Archambeau, et al.. How to assess FAIRness to improve crediting and rewarding processes for data sharing? A step forward towards an extensive assessment grid. RDA 13th (P13) Plenary Meeting, Apr 2019, Philadelphia, United States. ⟨https://www.rd-alliance.org/rda-13th-plenary-meeting-information⟩. ⟨10.5281/zenodo.2625721⟩. ⟨hal-02094678⟩

Useful resources

ARK:https://tools.ietf.org/html/draft-kunze-ark-18B2HANDLE: https://github.com/EUDAT-B2SAFE/B2HANDLECROP Ontology:
https://github.com/bioversity/Crop-OntologyDOI: https://www.doi.org/ePIC: https://www.pidconsortium.eu/GUID:
https://fr.wikipedia.org/wiki/Globally_Unique_IdentifierHandle System Namespace and Service Definition;
http://www.ietf.org/rfc/rfc3651.txtHandle System Protocol (ver 2.1) Specification http://www.ietf.org/rfc/rfc3652.txtHTML:
https://www.w3.org/html/
HTTP protocol: https://www.w3.org/Protocols/IRI: https://www.ietf.org/rfc/rfc3987.txt
Linked data: https://www.w3.org/wiki/LinkedDataLSID: http://www.lsid.info/
ORCID: https://orcid.org/

OWL: https://www.w3.org/OWL/Plant Trait Ontology:
https://github.com/Planteome/plant-trait-ontologyPURL:
https://en.wikipedia.org/wiki/Persistent_uniform_resource_locatorRDF
(W3C): https://www.w3.org/RDF/RDF-S:
https://www.w3.org/TR/rdf-schema/SKOS:
https://www.w3.org/TR/skos-reference/SPARQL:
https://www.w3.org/TR/rdf-sparql-query/URI:
https://www.w3.org/wiki/URIURL: https://www.w3.org/TR/url/URN:
https://www.w3.org/urn/UUID: https://www.w3.org/wiki/UriSchemes/uuidXRI
(OASIS): https://www.oasis-open.org/committees/xri/