Fair Data Points on standardised Plant data within WUR

Let data find itself

12-04-2018, Patrick Hendrickx

Outline of the presentation

- How does real life data look like? (start situation)
- Why is data unfindable, even though it has been saved in a database?
- Let data find itself by Linked Data.
- Example how it could work.

Data is modified for this example

The life of a researcher

- Produce a large volume high quality Data.
- Answer the research question.
- Write a report.
- Spend more hours than budgeted……
- Make the customer happy.
- Go on with the other projects because they have to be ready at the end of the year……

Why invest time in making data FAIR?????

And then we end up with a dataset

Accession	trait	score
PH001	Total Yield	5
PH002	Total Yield	7

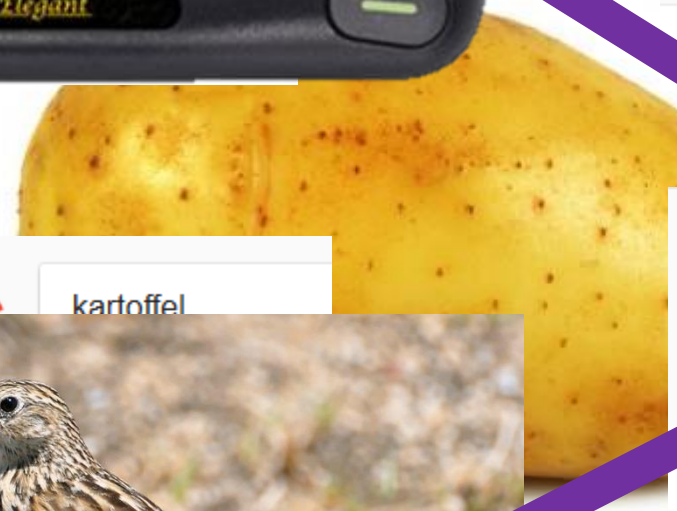
- When did they stop harvesting?
- What is the quality of the fruits?
- Is this value a mean of a plot or is it a plant?
- What is the source of the accession?
- What is the unit of the score 5 kg or (1=bad, 10=good)?
- Cultivation system, climate, watering, pruning?
- Can you access and find the dataset after 10 years?

Spend more energy in describing the datasets !!!!

- Design experiments in a FAIR point of view.
- Use wide used standards.
- Use well described methods or describe them yourself.

If you don't.....People can't find enough information about your hard work!!!!!

Lets search data about potato?



Google aardappel

Alle Afbeeldinge

Ongeveer 5.800.000

Google kartoffel

Google pieper

Alle Afbeeldingen S

Ongeveer 8.160.000 resultate

WAGENINGEN
UNIVERSITY & RESEARCH

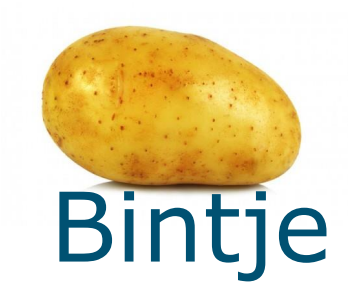
100years
1918 — 2018

Let experts create that universal language

- How do we call things (objects).
- How are the objects related with each other.
- What restrictions do we have.

We call that an ontology

How does an ontology solve the potato problem.



OLS > NCBI organismal classification NCBITAXON > NCBITaxon:4113

Solanum tuberosum

http://purl.obolibrary.org/obo/NCBITaxon_4113

has exact synonym
potatoes, potato

has obo namespace
ncbi_taxonomy

has rank
http://purl.obolibrary.org/obo/NCBITaxon_species

has related synonym
Solanum tuberosum subsp. tuberosum

owl:sameAs :

- dbpedia-cs:Lilek_brambor
- dbpedia-de:Kartoffel
- dbpedia-el:Πατάτα
- dbpedia-es:Solanum_tuberosum
- dbpedia-fr:Pomme_de_terre
- dbpedia-it:Solanum_tuberosum
- dbpedia-ja:ジャガイモ
- dbpedia-ko:감자
- dbpedia-nl:Aardappel
- dbpedia-pl:Ziemniak
- dbpedia-pt:Batata
- dbpedia-ru:Картофель

And there is a link between them!!!!

- Bintje —*Has Species Name*→ Solanum tuberosum

Our Job

- Solanum tuberosum —*GO:Has Exact Synonym*→ Potato
- Potato —*DBPedia:SameAs*→ aardappel

Work done by others

Computer readable format

- Solanum tuberosum —*GO:Has Exact Synonym*→ potato

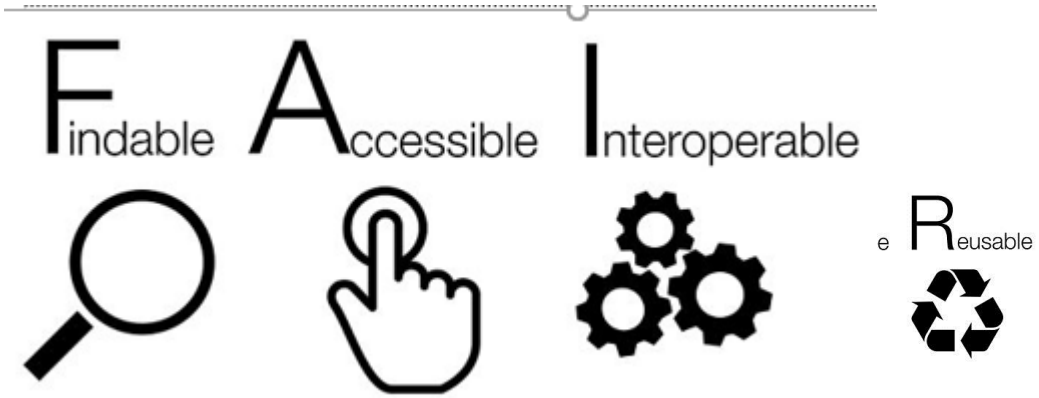
http://purl.obolibrary.org/obo/NCBITaxon_4113

http://www.geneontology.org/formats/oboInOwl#hasExactSynonym

http://purl.obolibrary.org/obo/po#Potato

Make data readable for computers.

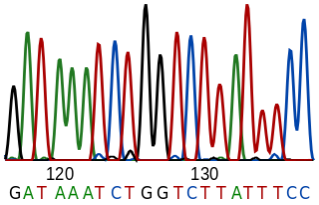
- Computers are much faster in searching for relevant data.
- Standardised protocols make it easier to find relevant data produced by third parties.
- If the computer can find data and use it.....



Example tomato

Does the country of origin have an effect on the fruit size of tomato?

- 105 completely different tomato accessions where sequenced.
- Fruit size is explained by multiple genes. Let's have a look at one of them.
- The program Haplosmasher (Plant Breeding) shows that there are 13 variants of this one gene.
- The 105 accessions are grouped by these variants.



105 sequence files

Haplosmasher

Gene V01

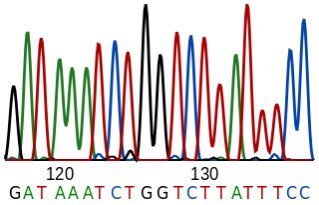
Gene V02

Grouped by haplotype:

.....	LA1044 - S. galapagense	RF_303 - -	RF_234 - S. lycopersicum cv 981136	LYC3476 - S. lycopersicum cv Lidi	RF_007 - S. lycopersicum cv Katinka Cherry	LYC2910 - S. pimpinellifolium (Jusl.) Mill.	RF_093 - S. lycopersicum cv Kentucky Beefsteak	PI169588 - S. lycopersicum cv Dolmalik	PI203232 - S. lycopersicum cv Wheatley_s Frost Resistant	RF_026 - S. lycopersicum cv Polish Joe	RF_310 - -
.|s1.....	CGN15464 - S. lycopersicum cv Rote Beere	LA1421 - S. lycopersicum cv	LA1324 - S. lycopersicum								
s1.....	RF_301 - -	RF_237 - S. lycopersicum var cerasiforme	LYC2962 - S. lycopersicum var. ventura	CGN15820 - S. lycopersicum	LA1718 - S. habrochaites	RF_226 - S. lycopersicum cv DL/67/248	LA1479 - S. lycopersicum var cerasiforme	LA1578 - S. pimpinellifolium	LA4451 - S. lycopersicum cv Black Cherry	RF_238 - S. lycopersicum cv R226	LYC2740 - S. pimpinellifolium
s1.....|s1.|s1.|s1.s2...	LA1954 - S. peruvianum										
s1.....|s1.|s1.|s1.|s1.|s1...	T1248 - S. corneliomulleri	CGN15530 - S. chilense									
s1.....|s1..|s1.|s1.|s1.|s1...	LA1983 - S. huaylasense										
s1.....|s1.....	CGN15791 - S. habrochaites f. glabratum	CGN15792 - S. habrochaites f. glabratum	LA1777 - S. habrochaites								
s1.....|s1...|s1.|s1..|s1...	LA2663 - S. chiemliewskii	LA2172 - S. arcanum	LA2133 - S. neorickii	LA2157 - S. arcanum							
s1.....|s1...|s1.|s1..|s1...|s1	LA2695 - S. chiemliewskii										
s1.....|s1..|s1.|s1.|s1..|s1...	LA1278 - S. peruvianum										
s1.....|s1..|s1.|s1.|s1..|s1..|s1.	LYC1831 - S. pennellii										
s1.....|s1...|s1..|s1.|s1.|s1..|s1..|s1.	LA1365 - S. huaylasense										
s1.....|s1....|s1.....	LA0407 - S. habrochaites	LYC4 - S. habrochaites									

How do I know which haplotypes corresponds with the big tomato's

- Plant Breeding has already some experimental data available in linked format.
- The standard used is MIAPPE (Minimal Information About Plant Phenotyping Experiments).
- The MIAPPE ontology covers from observation in the field or lab, experiment, growth facility, environment, literature and many more.
- Not all the details about the data are recoverable so this data is not completely reusable.



105 sequence files

Haplosmasher

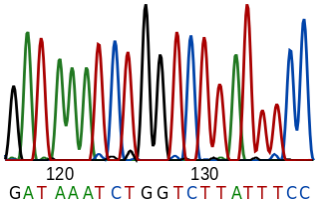
Grouped by haplotype:

.....	LA1044 - S. galapagense	RF_303 - -	RF_234 - S. lycopersicum cv 981136	LYC3476 - S. lycopersicum cv Lidi	RF_007 - S. lycopersicum cv Katinka Cherry	LYC2910 - S. pimpinellifolium (Jusl.) Mill.	RF_093 - S. lycopersicum cv Kentucky Beefsteak	PI169588 - S. lycopersicum cv Dolma	PI203232 - S. lycopersicum cv Wheatley_s Frost Resistant	RF_026 - S. lycopersicum cv Polish Joe	RF_310 - -
.|s1.....	CGN15464 - S. lycopersicum cv Rote Beere	LA1421 - S. lycopersicum cv	LA1324 - S. lycopersicum								
s1.....	RF_301 - -	RF_237 - S. lycopersicum var cerasiforme	LYC2962 - S. lycopersicum var. Ventura	CGN15820 - S. lycopersicum	LA1718 - S. habrochaites	RF_226 - S. lycopersicum cv DL/67/248	LA... - S. ...	LA1578 - S. pimpinellifolium	LA4451 - S. lycopersicum cv Black Cherry	RF_238 - S. lycopersicum cv R226	LYC2740 - S. pimpinellifolium
s1.....|s1.|s1.|s1.s2...	LA1954 - S. peruvianum										
s1.....|s1.|s1.|s1.|s1.|s1...	T1248 - S. corneliomulleri	CGN15530 - S. chilense									
s1.....|s1..|s1.|s1.|s1.|s1...	LA1983 - S. huaylasense										
s1.....|s1.....	CGN15791 - S. habrochaites f. glabratum	CGN15792 - S. habrochaites f. glabratum	LA1777 - S. habrochaites								
s1.....|s1...|s1.|s1..|s1...	LA2663 - S. chiemliewskii	LA2172 - S. arcanum	LA2133 - S. neorickii	LA2157 - S. arcanu...							
s1..... ...|s1...|s1	LA2695 - S. chiemliewskii										
s1..... ...|s1...	LA1278 - S. peruvianum										
s1..... ...|s1..|s1.	LYC1831 - S. pennellii										
s1..... ...|s1..|s1..|s1.	LA1365 - S. huaylasense										
s1.....|s1....|s1.....	LA0407 - S. habrochaites	LYC4 - S. habrochaites									

Example data

How do we get the country of origin?

- Gene banks are the most reliable source of data.
- Centre of Genetic Resources in Wageningen spent a lot of time in curating data.
- For this user case we converted the CGN tomato passport data to linked data with two standards.
    - FAO/Multi Crop Passport Descriptors (MCPD)
    - Minimum Information About Plant Phenotyping Experiment (MIAPPE)



105 sequence files

Haplosmasher

Grouped by haplotype:

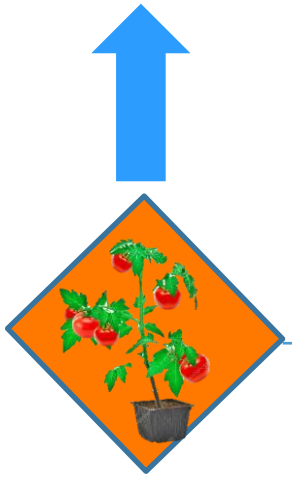
.....	LA ga	RF_303 - -		LYC3476 - S. lycopersicum cv Lidi	RF_007 - S. lycopersicum cv Katinka Cherry	LYC2910 - S. pimpinellifolium (Jusl.) Mill.		PI169588 - S. lycope cv Dolma	PI203232 - S. lycopersicum cv Wheatley_s Frost Resistant	RF_026 - S. lycopersicum cv Polish Joe	
.|s1.....	CGN15464 - S. lycopersicum cv Rote Beere	LA1421 - S. lycopersicum cv	LA1324 - S. lycopersicum								
s1..... .	RF_301 - -	RF_237 - S. lycopersicum var cerasiforme	LYC2962 - S. lycopersicum var. Ventura		LA1718 - S. habrochaites	RF_226 - S. lycopersicum cv DL/67/248			LA4451 - S. lycopersicum cv Black Cherry	RF_238 - S. lycopersicum cv R226	LYC2740 - S. pimpinellifolium
s1.....|s1.|s1.|s1.s2...	LA1954 - S. peruvianum										
s1.....|s1.|s1.|s1.|s1.|s1...	T1248 - S. corneliomulleri	CGN15530 - S. chilense									
s1.....|s1..|s1.|s1.|s1.|s1...	LA1983 - S. huaylasense										
s1.....|s1.....	CGN15791 - S. habrochaites f. glabratum		LA1777 - S. habrochaites								
s1.....|s1...|s1.|s1..|s1...	LA2663 - S. chiemliewskii	LA2172 - S. arcanum	LA2133 - S. neorickii	LA2157 - S. arcanu							
s1..... |s1...|s1	LA2695 - S. chiemliewskii										
s1..... ..|s1...											
s1..... ..|s1..|s1.											
s1..... |s1..|s1..|s1.	LA1365 - S. huaylasense										
s1.....|s1....|s1.....	LA0407 - S. habrochaites	LYC ha									

Example data

WAGENINGEN UNIVERSITY & RESEARCH

100years 1918 — 2018

Haplosmasher



Has tomato size

FDP:
WUR Plant
Breeding

Tomato phenotype data

has country of origin

FDP:
WUR CGN

Passport data

More linked data is needed!

- Data about genes, proteins and species are well covered.
- Missing is passport data and phenotype data.
- This data should be generated by researchers.

BUT

- Transfer data to linked data is not easy.
- Tools should make life easier.

Special thanks to,

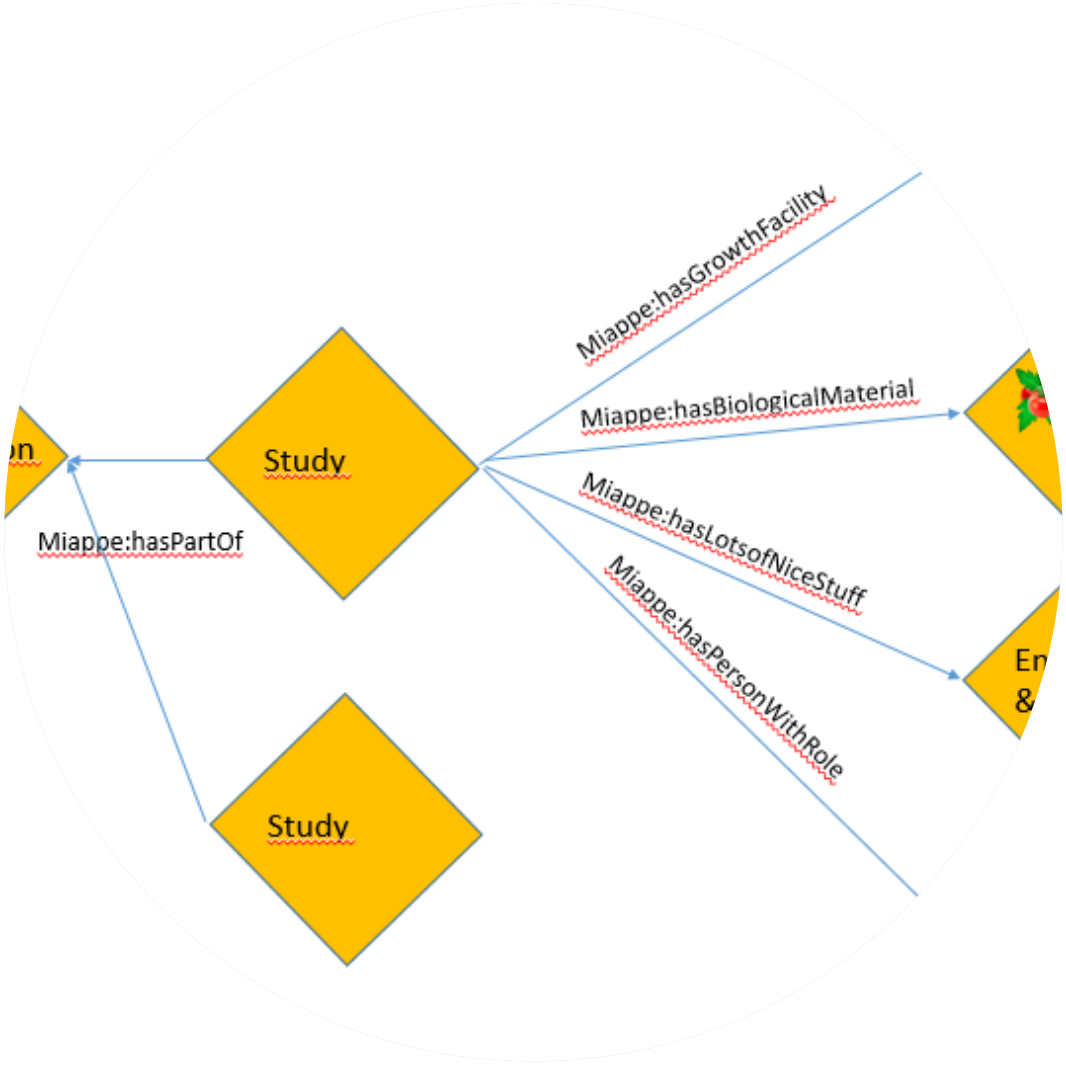
Eliana Papoutsoglou

Martijn van Kaauwen

Richard Finkers

CGN Wageningen

MIAPPE consortium

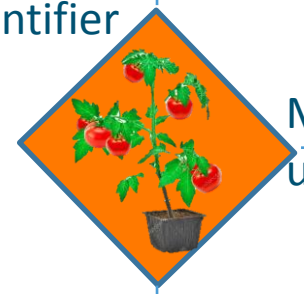


Multi Crop Passport Descriptors and
MIAPPE (Minimal Information About Plant Phenotyping Experiments)

CGN
1300 Accessions

BreeDB
8800 Accessions

Genesys PGR
38.000 Accessions



Miappe:hasTaxonIdentifier

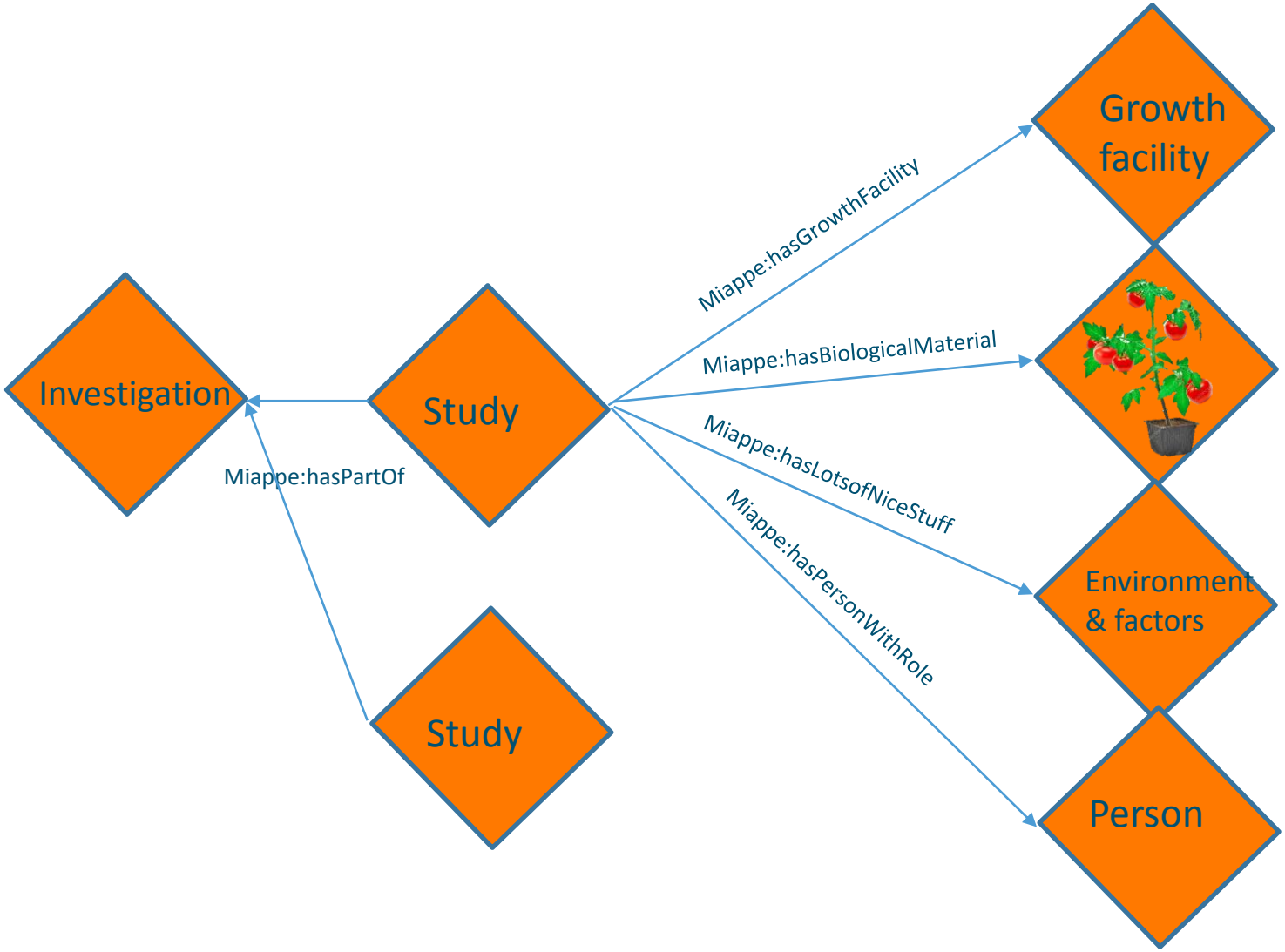
Miappe:hasDonorInstitute

Miappe:has country of origin

NCBI

DBpedia

DBpedia



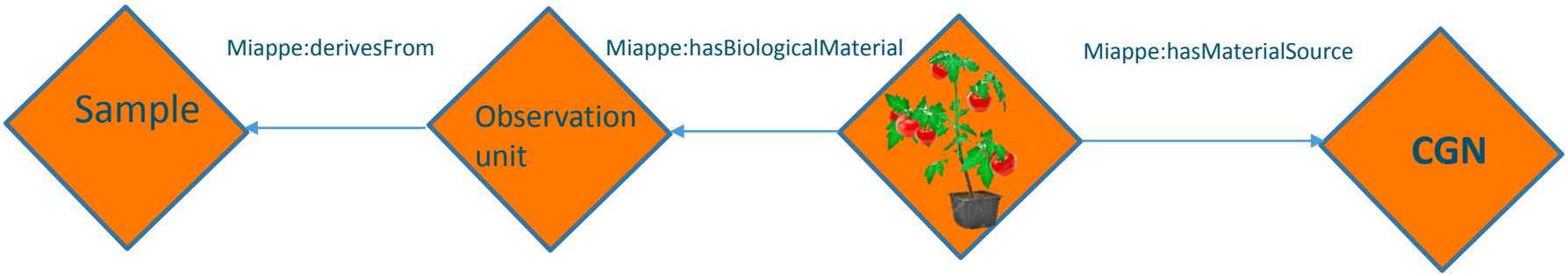
- Field
- Greenhouse

- Genotype
- Accession

WAGENINGEN
UNIVERSITY & RESEARCH

100years
1918 — 2018

Biological Material



- DNA
- HPLC
- Fruit
- Leaf

- Plot
- Plant

- Genotype
- Accession

- Accession

BreeDB

Database Genetic Resource Centre

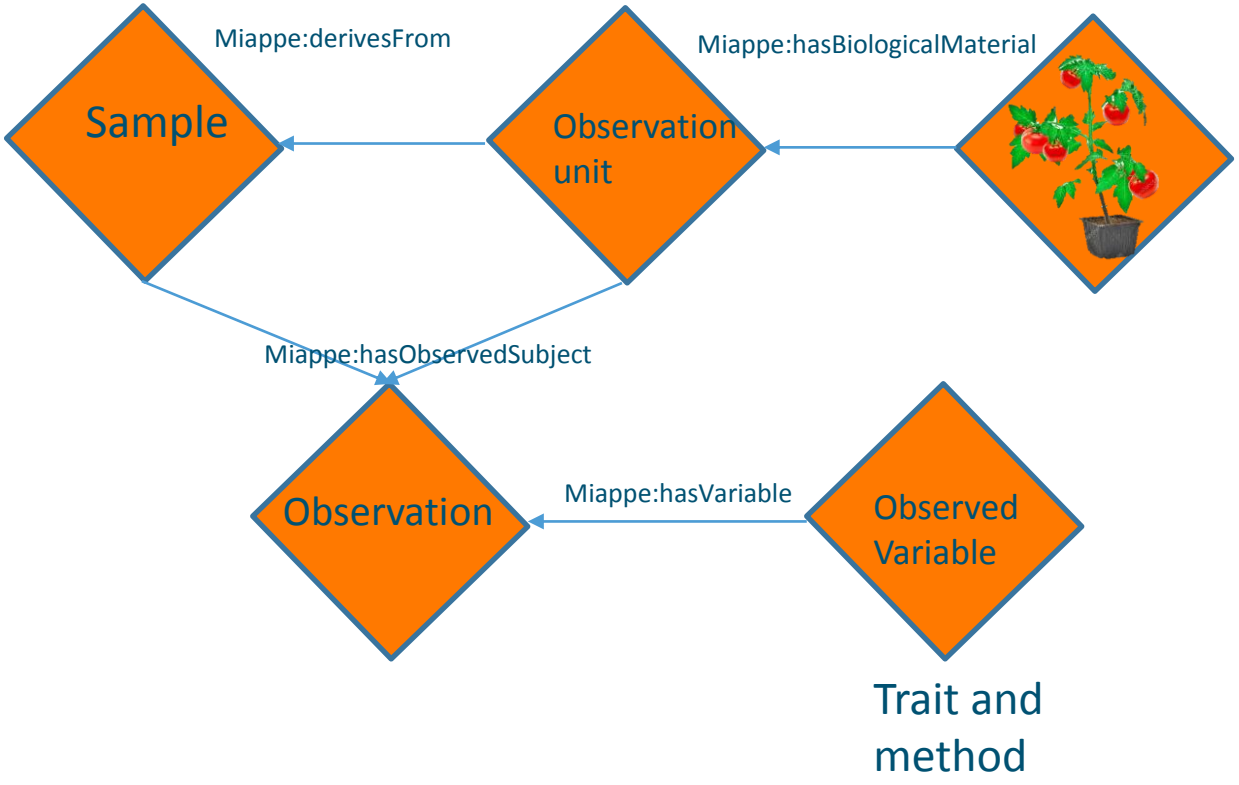
LabData

BreeDB

WAGENINGEN
UNIVERSITY & RESEARCH

100years
1918 — 2018

Observation



BreeDB MIAPPE compliant (experiments)

- Start Date and End Date missing or not available.
- Statistical design needs more options.
- Growth facility description is needed. Sometimes is this described in the experimentDescription.
- Project description is needed.
- Climate and Environmental parameters.

BreeDB MIAPPE compliant (observations)

- Method description is not sufficient.
- Observations not standardised.
- Observed Object is not well described.
  - Belongs the value to a single fruit or is it an average of fruits (accession)?
- No units or scales.
- Data is raw. Not curated.