BigDataGrapes WP3 Status

Vladimir Alexiev, PhD, PMP & Milena Yankova, PhD

2018-07-18, Montpellier

ontotext

Overview

- Achievements
- Ontology Notes
- WP3 Problems and Steps
- Long-Term Plans

Achievements

Scope and Meetings

- WP3 Data & Semantics Layer is a core WP of the project
  - If we have no data, we can have no achievements
- Between 30 Apr and 16 July, ONTO hosted 6 project meetings

- Currently working on T3.1 Data Modelling over Big Data Infrastructures (**ONTO, AGRO, AUA, INRA**)
- After we figure out a semantic data model, we should start on T3.2 Data Ingestion and Integration
- In parallel (and very soon!), we should start on T3.3 Big Data Indexing (also **CNR**)
- Participants (most active listed first):
  - ONTO: Vladimir
  - AGRO: Antonis, Pythagoras
  - AUA: Katerini
  - INRA: Sabine-Karen, Danai?
  - Occasional: CNR, Geocledian, Apigea

Semantic Data Integration Steps

- data analysis
- defining data requirements (competence questions)
- ontology engineering: selection, combination and extension of ontologies
- *(we are somewhere here)*
- semantic modeling and creating application profiles and/or RDF shapes (SHACL, ShEx)
- semantic conversion and tools, depending on source (CSV/TSV tabular, RDBMS, XML)
- semantic alignment and instance matching
- URL design, semantic publishing, content negotiation
- data validation and data quality management/measurement
- data update flows
- model documentation, sample queries, deploying queries as services

Minutes and Resources

- gfolder WP3
- gdoc WP3-meeting-minutes has: minutes, background from DOW, links to resources, detailed material

- When some material gets bigger, we split it off to a separate file
- Github repo: https://github.com/BigDataGrapes-EU/ontology (first WP in the project)
  - data: semantic data (for now, some samples)
  - ttl: relevant ontologies converted to turtle (and added prefixes) for easier reading
  - misc: ontology materials in miscellaneous formats (eg xlsx, obo)
  - notes: various notes on ontologies and data.
- Mirroring between gfolder and Github, bad idea, **must decide** Github or Google Drive

Ontologies Researched

- AGRO submitted sheet Relevant Ontologies and Vocabularies: 16 relevant ontologies, plus 21 on specific crops to use as examples
- ONTO downloaded 17 ontologies, converted to ttl which is easier to read
- ONTO researched them to some extent and wrote up various problems (see next)
- ONTO also researched several ontology portals, with a total of 200 ontologies, 5M classes (!), 16k props, 476k individuals

Ontology List

- AEO (OAE): Agricultural Experiments Ontology
- AFEO: Agri-Food Experiment Ontology
- AGRO: Agronomy Ontology
- AT: Agricultural Technology Ontology
- BCO: Biological Collection Ontology
- BFO: Basic Formal Ontology
- ChEBI: Chemical Entities of Biological Interest
- CO: Crop Ontology (series of)
  - CO_320: Rice
  - CO_322: Maize
  - CO_356: Vitis (viticulture)
  - CO_357: Woody Plant
  - CO_UO: Units Ontology
- EO (ENVO): Environment Ontology
- IAO: Information Artifact Ontology
- MMO: Measurement Methods Ontology
- NCBITaxon: NCBI Taxonomy
- OBO: Open Biological and Biomedical Ontology (a big set)
- OEPO: Ontology for Experimental Phenotypic Objects
- OFPE: Ontology for Food Processing Experiment
- PATO: Phenotypic Quality Ontology
- PCO: Population and Community Ontology
- PECO: Plant and Environemental Conditions Ontology
- PO: Plant Ontology
- RO: Relations Ontology
- SDGIO: SDG-Interface Ontology
- TO: Trait Ontology
- UO: Unit Ontology

Ontology Metrics

	Classes	Properties	Individuals
AEO	56	36	30
AFEO	68	8	0
AGRO	1685	709	284
BCO	157	279	28
BFO	35	20	
CHEBI	128900	45	
CO_356	814	10	
ENVO	8510	241	21
FOODON	27050	130	359
IAO	219	111	23
NCBITaxon	1692930	27	
...	...	...	...
TOTAL	1874943	3598	1467

Ontology Portals

Useful to search for terms, see total size, inspiration for our own tools

- OBOfoundry: list of ontologies, with resource links
- CropOntology: collaborative ontology development
  - CO Annotation: **annotate tabular data with terms**
- Planteome: PO, TO, EO. Tree browser, graph vis
- EBI OLS: 200 ontologies eg agro, Crop Ontologies, Tree browser, graph vis, useful search eg NDVI
  - EBI OXO: **Ontology Xref Service**: serves ontology mappings, will integrate to OLS in 2018
- GODAN VEST: AgriSemantics Map of Data Standards: 398 ontologies, 215 Food and agriculture, 76 from AgroPortal, 328 from VEST Registry. Eg AEO
- OntoBee: eg PO, AGRO. Detailed Statis, eg PO
- ABER OWL: eg AGRO. Simpler browsing

AUA Tabular Data

Table Grapes Data

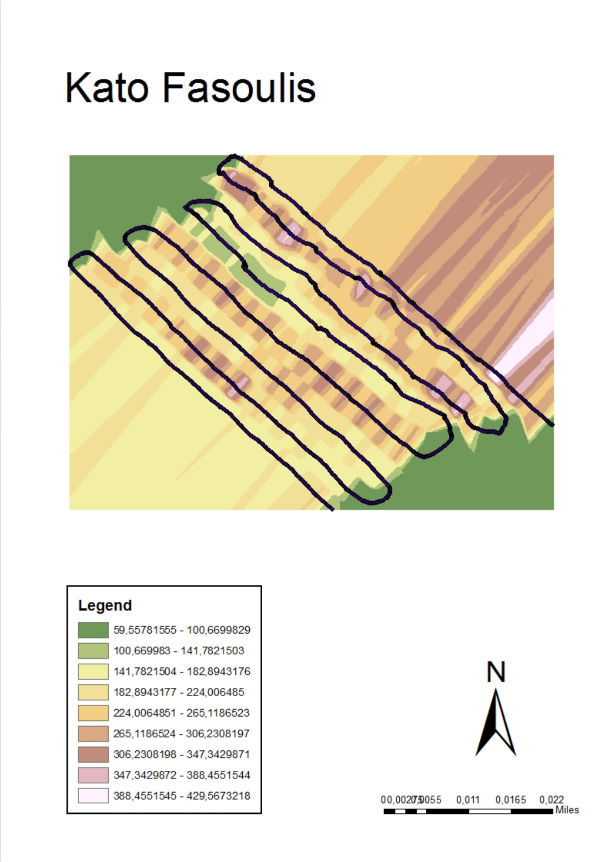
- Look at data in WP8/Table Grapes Pilot- AUA/Data

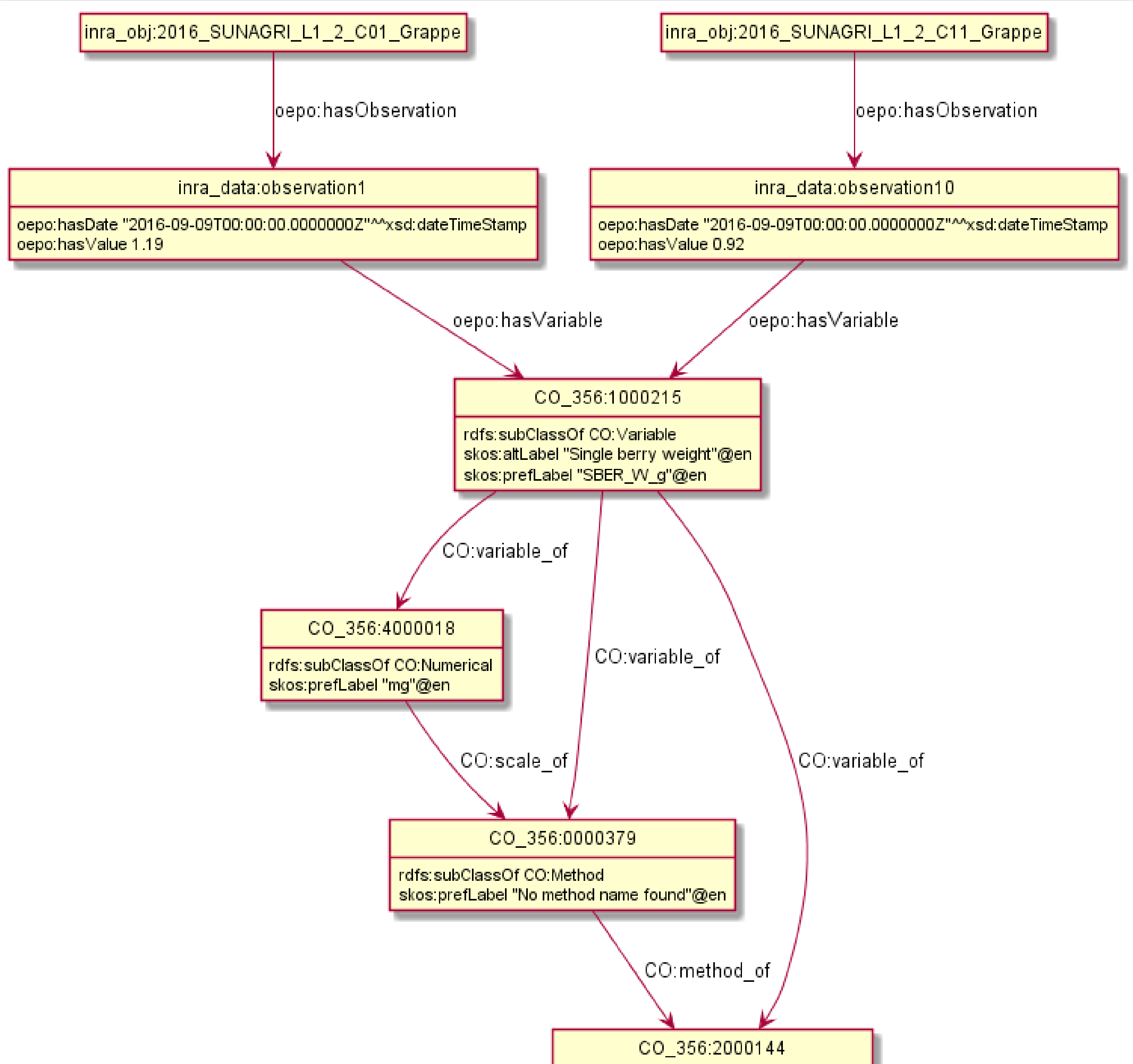
- Look in D8.1 Piloting Plan (specifically BigDataGrapes_Piloting Plan-AUA) for descriptions of equipment and measured indicators
- Tabular observation data: soil, plant canopies, spectral vegetation indexes
- 3 estates: Fasoulis, Kontogiannis, Palivou (see Photos for some images)
- Equipment: EM38, RapidScan, SpectroSense, Crop Circle
- Geo-referenced: longitude, latitude, altitude; timestamped
- About 10 measurements per measurement spot

Represents **realistic measurement data** - Currently working out a semantic representation to use as etalon for other data

Data Handling Needs

- To tie measurements to a specific plot: localization of geo-coordinates within the plot (GeoSPARQL predicate `within`)
- Discretization (then Averaging) to correlate measurements from different equipment and different days:
  - Of Geo-coordinates (eg to a grid of 2x2m), then
  - Of Datetime (eg is it ok to correlate two measurements done within a day? How about within a week?)
- Data cleaning: discard defective or outlier measurements
  - Eg1: RapidScan needs some “warm up” time to establish a GPS connection. Discard:
    - Readings with “FIXTYPE: Fix not valid” (missing)
    - Readings with negative ELEVATION (invalid)
  - Eg2: EM38 is affected by metal pillars, so conductivity readings above 100 should be discarded.
    - Eg on [Fasoulis_Kato_EM38_map](), only the green readings should be retained.
- Matches [Use case A. Data Anomaly Detection & Classification]().
  - Eg Eca sensing: Georeferenced soil electrical conductivity data; Operations: Data filtering for outliers





inra_obj:2016_SUNAGRI_L1_2_C01_Grappe

oepo:hasObservation

inra_data:observation1

oepo:hasDate "2016-09-09T00:00:00.0000000Z"^^xsd:dateTimeStamp
oepo:hasValue 1.19

inra_obj:2016_SUNAGRI_L1_2_C11_Grappe

oepo:hasObservation

inra_data:observation10

oepo:hasDate "2016-09-09T00:00:00.0000000Z"^^xsd:dateTimeStamp
oepo:hasValue 0.92

oepo:hasVariable

oepo:hasVariable

CO_356:1000215

rdfs:subClassOf CO:Variable
skos:altLabel "Single berry weight"@en
skos:prefLabel "SBER_W_g"@en

CO:variable_of

CO_356:4000018

rdfs:subClassOf CO:Numerical
skos:prefLabel "mg"@en

CO:variable_of

CO:variable_of

CO:scale_of

CO_356:0000379

rdfs:subClassOf CO:Method
skos:prefLabel "No method name found"@en

CO:method_of

CO_356:2000144

INRA RDF Data 3, 4 Feedback

Feedback ontology/notes#INRA Samples

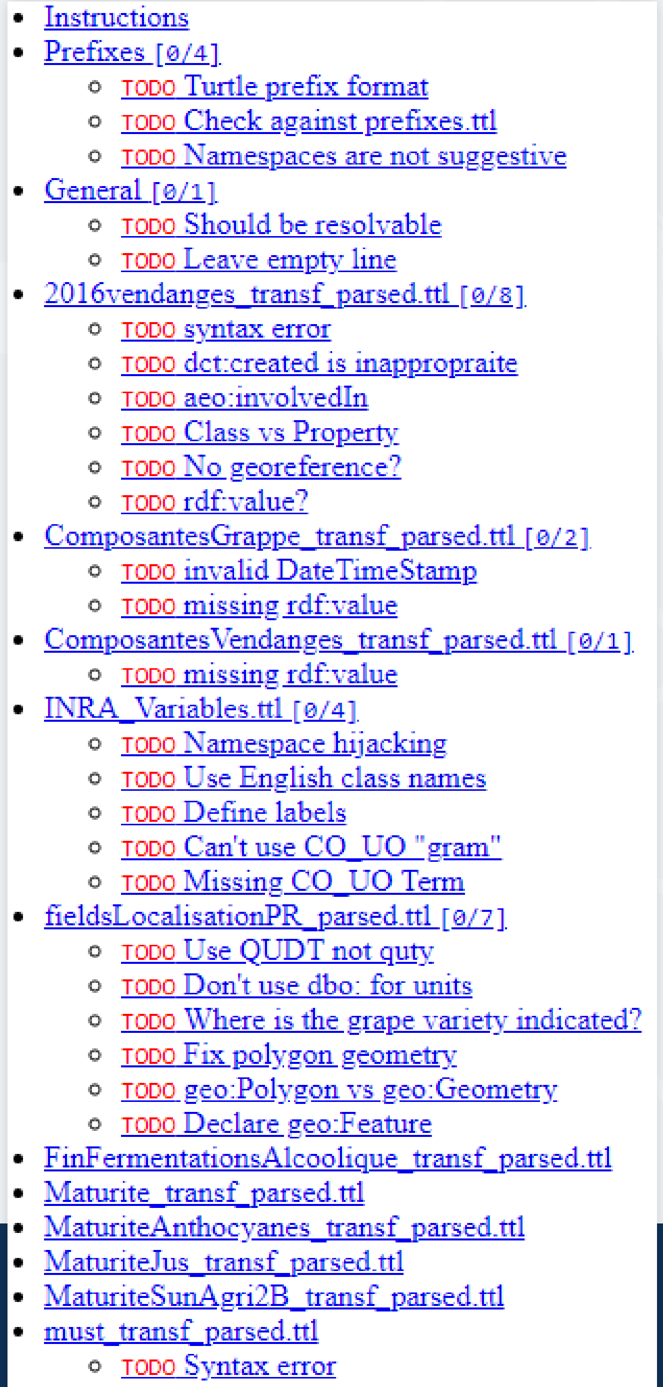
- Wrong URL http://www.cropontology.org/ontology/CO_356/Vitis#1000215, should be http://www.cropontology.org/rdf/CO_356:1000215
- `"2016-09-09T00:00:00.0000000Z"^^xsd:date` pads with fake time 0, uses invalid datatype (should be `xsd:dateTimeStamp`)
- Observations eg http://vinnotec.supagro.inra.fr/public/Pr/data/observation1 are missing `rdf:type`
- Observation objects eg http://vinnotec.supagro.inra.fr/public/Pr/2016_SUNAGRI_L1_2_C01_Grappe are not defined in these files

data/INRA/data5.

- Plot geometry using GeoSPARQL (`geo:asWKT`)
- Observations on harvest, fermentation, maturity, must, total sugars (BRIX refractometry)
- `INRA_Variables.ttl` defines extra terms (maybe they belong in Vitis?)

Fedback `README.org` and HTML rendered

- Detailed feedback provided on 7 files, still need to check 5 files
- About 30 TODO reported, must turn all of them into DONE



- Instructions
- Prefixes [0/4]
  - TODO Turtle prefix format
  - TODO Check against prefixes.ttl
  - TODO Namespaces are not suggestive
- General [0/1]
  - TODO Should be resolvable
  - TODO Leave empty line
- 2016vendanges_transf_parsed.ttl [0/8]
  - TODO syntax error
  - TODO dct:created is inappropraite
  - TODO aeo:involvedIn
  - TODO Class vs Property
  - TODO No georeference?
  - TODO rdf:value?
- ComposantesGrappe_transf_parsed.ttl [0/2]
  - TODO invalid DateTimeStamp
  - TODO missing rdf:value
- ComposantesVendanges_transf_parsed.ttl [0/1]
  - TODO missing rdf:value
- INRA_Variables.ttl [0/4]
  - TODO Namespace hijacking
  - TODO Use English class names
  - TODO Define labels
  - TODO Can't use CO_UO "gram"
  - TODO Missing CO_UO Term
- fieldsLocalisationPR_parsed.ttl [0/7]
  - TODO Use QUDT not quty
  - TODO Don't use dbo: for units
  - TODO Where is the grape variety indicated?
  - TODO Fix polygon geometry
  - TODO geo:Polygon vs geo:Geometry
  - TODO Declare geo:Feature
- FinFermentationsAlcoolique_transf_parsed.ttl
- Maturite_transf_parsed.ttl
- MaturiteAnthocyanes_transf_parsed.ttl
- MaturiteJus_transf_parsed.ttl
- MaturiteSunAgri2B_transf_parsed.ttl
- must_transf_parsed.ttl
  - TODO Syntax error

Data Handling and Validation

data-validation-handling. Started rules on:

- How to submit files (**must decide Github or gdrive**)
- How to use and update `prefixes.ttl`
- How to validate syntax using `riot` (and maybe `eyeball`)
- Once we decide on patterns for representing data, we will implement RDF Shapes for validating data requirements

Hope this will grow to a comprehensive doc on semantic data handling and validation by BDG project partners

Competence Questions

Competence Questions - What data you **have** - What data **needs** you have, or what questions the data should be able to answer

Given the abundance of available data and the sea of agro-bio ontologies, **data needs** are crucial to keep the modeling effort

focused, and drive these tasks:

- Seeking more data for specific questions
- Deciding which ontologies to involve and whether more ontological work is needed
- Structuring the data in an appropriate form (semantic modeling)
- Defining data tasks: conversion, cleanup/filtering, discretization…
- Creating sample queries to help data consumers

We need **real, validated** competence questions to drive our work

AGRO Competence Questions

Submitted draft Competence Questions:

1. Can I retrieve the sub-plots for a given plot?
2. Which varieties are cultivated in a given plot?
3. Can I retrieve weather data for a given plot?

4. Which varieties are cultivated in a soil with certain characteristics?
5. Can I retrieve the origin locale for a given test sample?
6. Can I retrieve images of a plot from which a sample was taken
7. Can I retrieve historical yield results for a plot (providing a timestamp)?
8. Can I retrieve historical weather data for a plot (providing a timestamp)?

Next steps

- ONTO started elaborating
- AGRO needs to add more details, down detalis such as
  - Are geolocation qualifiers (satellite, quality, HDOP) needed?
- All partners must validate questions and ground them in (derive them from) Use Cases

Data Domains

What data we need to represent?

- Observations: when (timestamp), where (georeference), what (measure, dimension, attribute, and observation)
- Estates and plots, including geospatial data
- Measurement equipment
- Experiments?
- Static nomenclature data, eg: varieties, types of measurement…
- Photos… of what?
- What else??

Ontology Notes

Ontology Problems

Github ontology/notes, rendered HTML

- Numerous prefix problems. Collected master file prefixes.ttl, let's use it
- Ontology namespace and ontology file differ significantly, no semantic resolution
- Some ontologies use wrong namespace (URLs don't mesh), eg Vitis uses `rdfs:subProperty` (it's `rdfs:subPropertyOf`)
- CO_357/nt is invalid, because someone was too lazy to put in new lines
- Various terms with unfilled labels, eg `CO_356:0000309` "name: No method name found"
- Viis Mismatch: is `CO_356:1000215` measured in *grams* (name "SBER_W_g") or *milligrams* (relation to `CO_356:4000018` "mg")
- Classes, properties and even some ontology files use numeric codes not English names
  - Makes it necessary to implement some search/browse interface to use them efectively
- Even `rdfs:label` often uses unreadable abbreviations
  - Eg `CO_322:0001093` "EWid_M_mm" means *Ear width, measurement, in mm*
  - Eg `CO_320:0000824` "PanLng_MatAv_UPOV1to3" means *rice panicle length, mature - average, UPOV scale, 1..3*
- Slash in local names make invalid prefixed names, eg `CO_322:0000320/2` is value 3="21-30% dead leaf area" of `CO_322:0000320` "0-10 Senescence scale"
- Space in URL is bad practice, eg `rice:Biotic%20stress`:
- Many terms declared both `skos:Concept, owl:NamedIndividual, owl:Class` and connected by both `rdfs:subClassOf` and `skos:broaderTransitive`.
  - represents heavy punning and makes OWL inference impossible

Over-Commitment

In many cases terms are defined at the wrong level of abstraction

- Eg NDVI is defined only in Maize, so it can't be used for grapes (Vitis)
  - Do we repeat the same mistake in Vitis, or try to move this to the Crop Ontology?
- Eg `CO_UO` defines "grams" relative to some woody plant feature. This is crazy because a gram is a gram, no matter what it's used to measure. So this unit cannot be used for grapes.

```
CO_UO:0000021 rdfs:label "g"@en; CO:scale_of CO_357:2000105.
CO_357:2000105 rdfs:label "Ratio shoot root protocol"@en
```

- It's better to use a proper Units ontology like QUDT, which defines units in terms of fundamentals (Mass, Length, Time, etc) and conversion factors between units
- Namespace hijacking: redefining imported (external) terms

Ontology Problems: Conclusion

- Property naming convention (lowerCamelCase) not followed, eg `po:Tomato rdfs:subPropertyOf oboInOwl:SubsetProperty`
- Improperly formatted timestamps, eg `"Jul 28, 2013 6:56:15 AM"^^xsd:dateTime`
- Wrong URL https://www.w3.org/TR/xmlschema-2/#rf-maxInclusive (in some text), should be http://www.w3.org/2001/XMLSchema#maxInclusive (semantic)
- Mixup of properties and URLs into a string, eg

```
oepo:WindSensor rdfs:isDefinedBy "skos:exactMatch http://purl.oclc.org/NET/ssnx/meteo/aws#WindSensor" ;
```

General conclusions

- Seems to me there's very little quality control in AgroBio ontologies
- Maybe a lot of these 200 ontologies and 5M terms are created just to do research, not used in real data
- Engage with the AgroBio community to fix some of the problems

WP3 Problems and Steps

Not Enough Traction

As you can see on slide “Semantic Data Integration Steps”, we’re still in the beginning. Reasons:

- No effective leadership by AGRO to collect competence questions and liaise with the AgroBio community
- Irregular meeting attendance: only one meeting was attended by all 4 organizations
- Little progress between meetings
- Feedback (error reports) provided by ONTO received no reaction
- No collaboration in Github yet (hopefully soon)

Immediate Next Steps

- AGRO: take the lead on validated Competence Questions
- ONTO & AGRO & INRA: establish Ontology Working Group: fixed responsibilities, regular meetings, progress between meetings
- AGRO & INRA: establish collaboration process with the AgroBio community (see 20180623 Modeling Quality): have direct contacts to the persons behind these ontologies (and/or the relevant curation teams)
- INRA & ONTO: map AUA tabular data to ontologies

Short-Term Plans

DONE

- Research ontologies sent by partners and other related ontologies (AGRO, INRA)

IN-PROGRESS

- Get competence questions (AGRO, all partners)
- Get sample tabular data from partners (AUA, others)
- Get sample RDF data from partners (INRA)
- Report ontology and instance data errors to partners

TODO

- Report ontology errors to AgroBio community and engage to fix them
- Discuss how to represent various Data Domains with partners
- Create a semantic model with rdfpuml
- Create text narrative (see euBusinessGraph Semantic Model as an example)
- Get the model approved

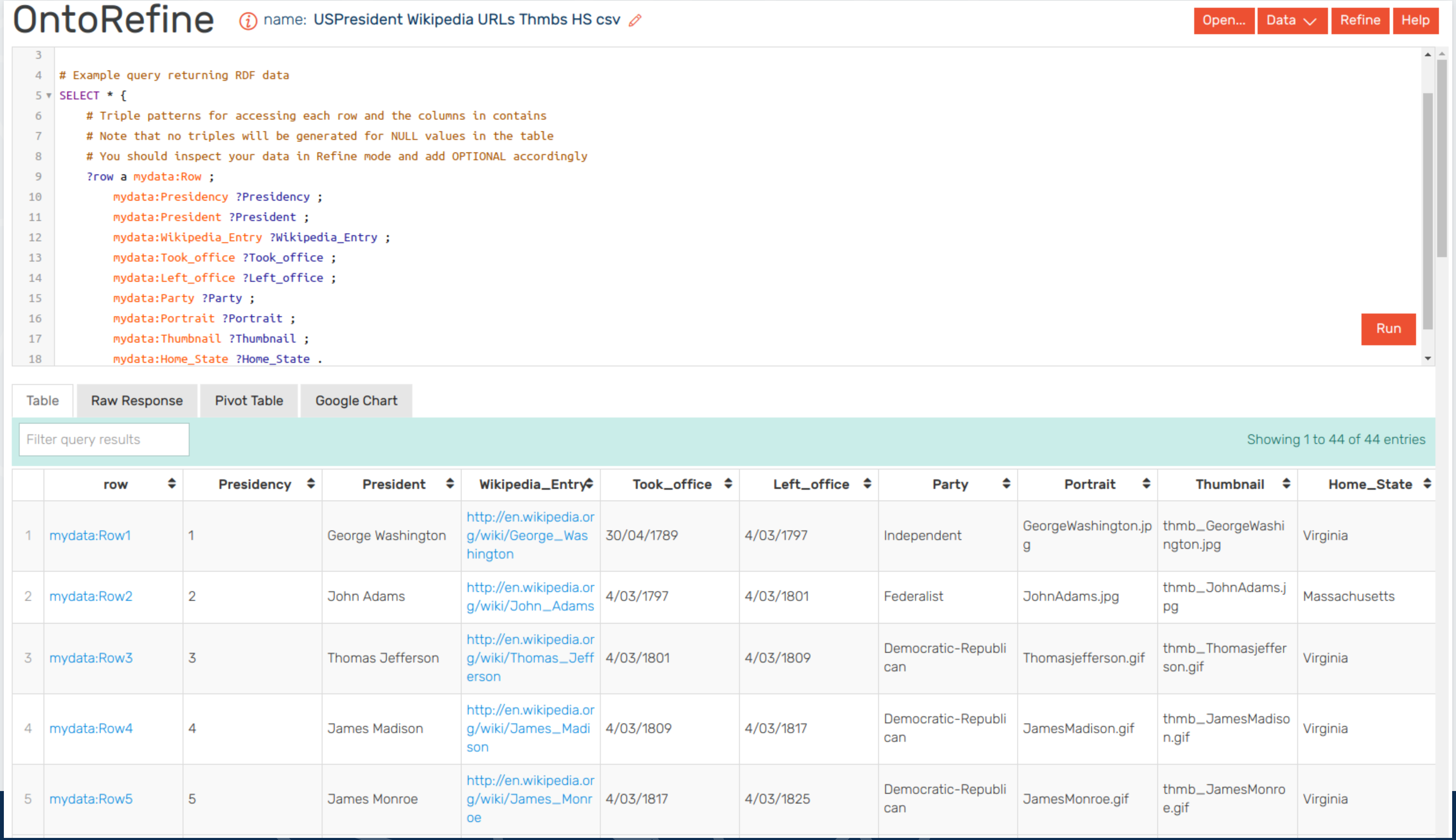
Long-Term Plans

Long-Term Plans

- Create RDF shapes for the approved model (SHACL and/or ShEx), establish validation
- Research and specify possible tools (conversion, annotation, search)
- Select or implement/deploy tools
- Implement special data processing (eg cleaning, discretization)
- Establish data ingestion pipeline
- Load data to GraphDB
- Implement sample queries

Possible Tool: GDB OntoRefine

OntoRefine: part of GDB Workbench, easy cleaning and conversion of tabular data.



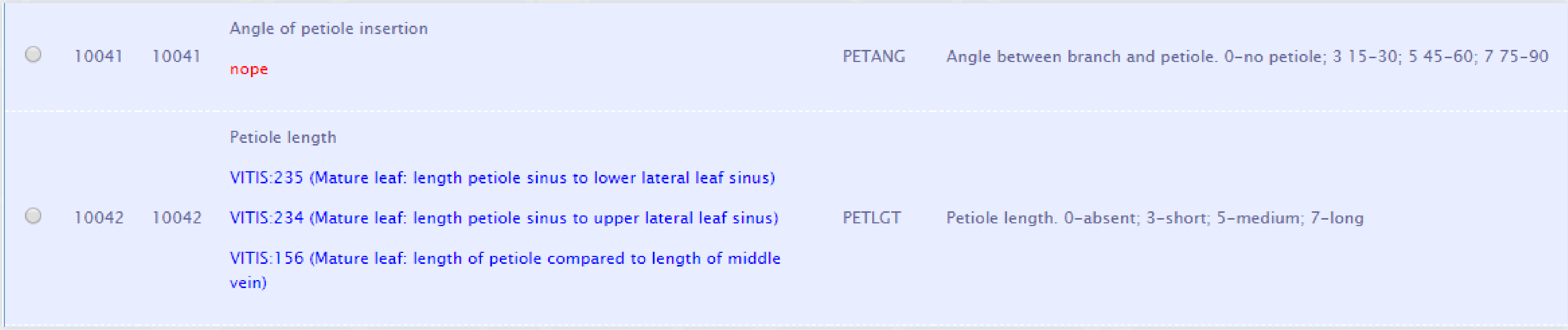
```
# Example query returning RDF data
SELECT * {
    # Triple patterns for accessing each row and the columns in contains
    # Note that no triples will be generated for NULL values in the table
    # You should inspect your data in Refine mode and add OPTIONAL accordingly
    ?row a mydata:Row ;
        mydata:Presidency ?Presidency ;
        mydata:President ?President ;
        mydata:Wikipedia_Entry ?Wikipedia_Entry ;
        mydata:Took_office ?Took_office ;
        mydata:Left_office ?Left_office ;
        mydata:Party ?Party ;
        mydata:Portrait ?Portrait ;
        mydata:Thumbnail ?Thumbnail ;
        mydata:Home_State ?Home_State .
```

Showing 1 to 44 of 44 entries

	row	Presidency	President	Wikipedia_Entry	Took_office	Left_office	Party	Portrait	Thumbnail	Home_State
1	mydata:Row1	1	George Washington	http://en.wikipedia.org/wiki/George_Washington	30/04/1789	4/03/1797	Independent	GeorgeWashington.jpg	thmb_GeorgeWashington.jpg	Virginia
2	mydata:Row2	2	John Adams	http://en.wikipedia.org/wiki/John_Adams	4/03/1797	4/03/1801	Federalist	JohnAdams.jpg	thmb_JohnAdams.jpg	Massachusetts
3	mydata:Row3	3	Thomas Jefferson	http://en.wikipedia.org/wiki/Thomas_Jefferson	4/03/1801	4/03/1809	Democratic-Republican	Thomasjefferson.gif	thmb_Thomasjefferson.gif	Virginia
4	mydata:Row4	4	James Madison	http://en.wikipedia.org/wiki/James_Madison	4/03/1809	4/03/1817	Democratic-Republican	JamesMadison.gif	thmb_JamesMadison.gif	Virginia
5	mydata:Row5	5	James Monroe	http://en.wikipedia.org/wiki/James_Monroe	4/03/1817	4/03/1825	Democratic-Republican	JamesMonroe.gif	thmb_JamesMonroe.gif	Virginia

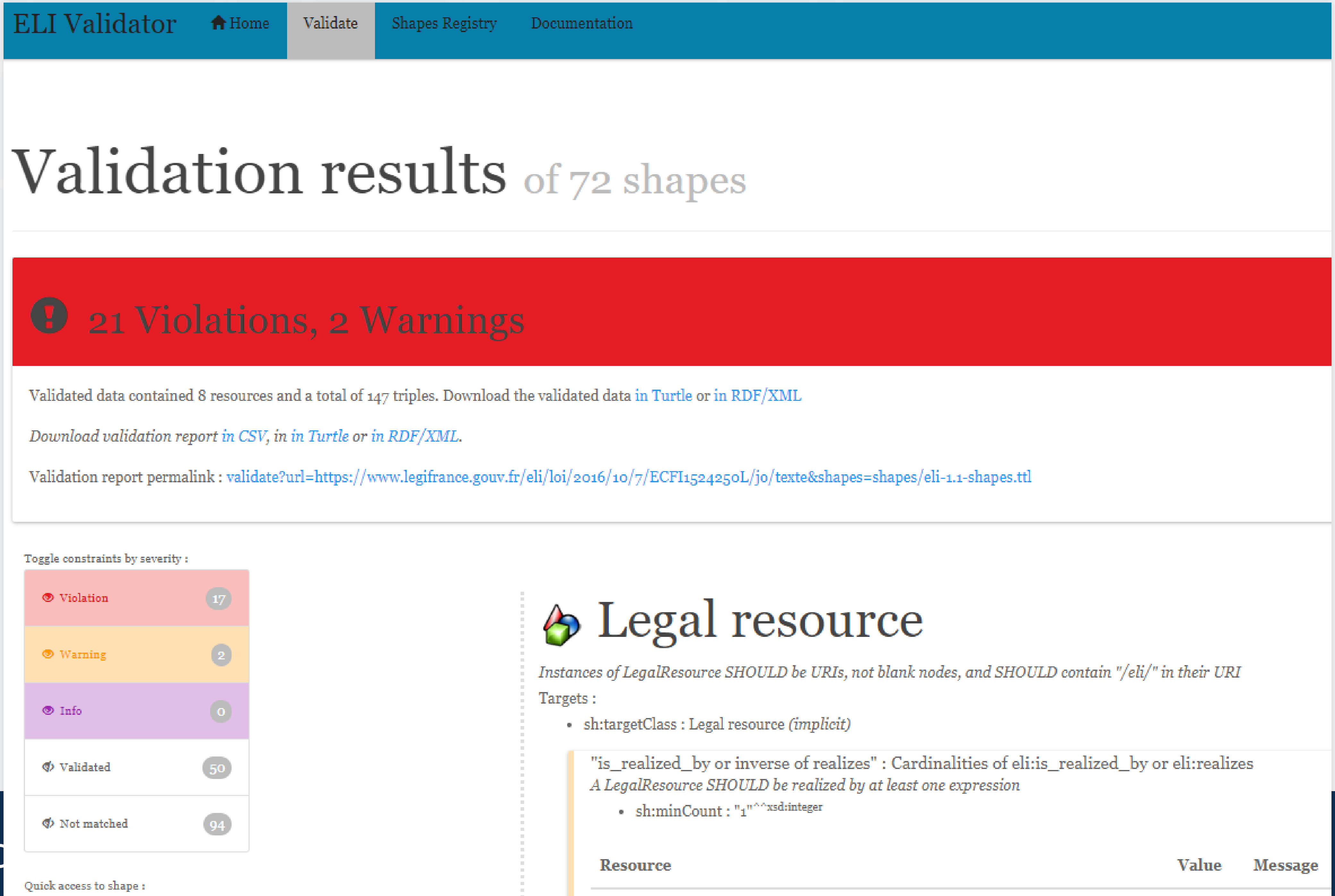
Possible Tool: CO Term Annotation

Eg CO Annotation: annotate tabular data with terms



○	10041	10041	Angle of petiole insertion nope	PETANG	Angle between branch and petiole. 0-no petiole; 3 15-30; 5 45-60; 7 75-90
○	10042	10042	Petiole length VITIS:235 (Mature leaf: length petiole sinus to lower lateral leaf sinus) VITIS:234 (Mature leaf: length petiole sinus to upper lateral leaf sinus) VITIS:156 (Mature leaf: length of petiole compared to length of middle vein)	PETLGT	Petiole length. 0-absent; 3-short; 5-medium; 7-long

Shows that many terms are not found, and ambiguity of terms between ontologies (over-specification)



Validation results of 72 shapes

❗ 21 Violations, 2 Warnings

Validated data contained 8 resources and a total of 147 triples. Download the validated data in Turtle or in RDF/XML

Download validation report in CSV, in in Turtle or in RDF/XML.

Validation report permalink : validate?url=https://www.legifrance.gouv.fr/eli/loi/2016/10/7/ECFI1524250L/jo/texte&shapes=shapes/eli-1.1-shapes.ttl

Toggle constraints by severity :

Violation	17
Warning	2
Info	0
Validated	50
Not matched	94

Quick access to shape :

Legal resource

Instances of LegalResource SHOULD be URIs, not blank nodes, and SHOULD contain "/eli/" in their URI

Targets :

- sh:targetClass : Legal resource *(implicit)*

"is_realized_by or inverse of realizes" : Cardinalities of eli:is_realized_by or eli:realizes
A LegalResource SHOULD be realized by at least one expression

- sh:minCount : "1"$^{\wedge\wedge xsd:integer}$

Resource	Value	Message

Possible Extensions/Integrations

- Create library of shapes (SHACL or ShEx) for validation, with visualization
- Add library of data shapes (SPARQL CONSTRUCT) to OntoRefine
- Integrate shape validation to OntoRefine

Etc etc etc. But we first need specifications!