Ref. Ares(2018)5105768 - 04/10/2018

Big Data to Enable Global Disruption of the Grapevine-powered Industries

D3.1 - Data Modelling and Linking Components

DELIVERABLE NUMBER	D3.1
DELIVERABLE TITLE	Data Modelling and Linking Components
RESPONSIBLE AUTHOR	Vladimir Alexiev (ONTOTEXT)

Co-funded by the Horizon 2020
Framework Programme of the European Union

GRANT AGREEMENT N.	780751
PROJECT ACRONYM	BigDataGrapes
PROJECT FULL NAME	Big Data to Enable Global Disruption of the Grapevine-powered industries
STARTING DATE (DUR.)	01/01/2018 (36 months)
ENDING DATE	31/12/2020
PROJECT WEBSITE	http://www.bigdatagrapes.eu/
COORDINATOR	Pythagoras Karampiperis
ADDRESS	110 Pentelis Str., Marousi, GR15126, Greece
REPLY TO	pythk@agroknow.com
PHONE	+30 210 6897 905
EU PROJECT OFFICER	Mr. Riku Leppanen
WORKPACKAGE N. | TITLE	WP3 | Data & Semantics Layer
WORKPACKAGE LEADER	ONTOTEXT
DELIVERABLE N. | TITLE	D3.1 | Data Modelling and Linking Components
RESPONSIBLE AUTHOR	Vladimir Alexiev (ONTOTEXT)
REPLY TO	vladimir.alexiev@ontotext.com
DOCUMENT URL	http://www.bigdatagrapes.eu/
DATE OF DELIVERY (CONTRACTUAL)	30 September 2018 (M9)
DATE OF DELIVERY (SUBMITTED)	4 October 2018 (M10)
VERSION | STATUS	1.0 | Final
NATURE	Demonstrator (DEM)
DISSEMINATION LEVEL	Public (PU)
AUTHORS (PARTNER)	Vladimir Alexiev (ONTOTEXT)
CONTRIBUTORS	Franco Maria Nardini (CNR), Raffaele Perego (CNR), Nicola Tonellotto (CNR), Pythagoras Karampiperis (Agroknow), Antonis Koukourikos (Agroknow)
REVIEWER	Franco Maria Nardini (CNR)

VERSION	MODIFICATION(S)	DATE	AUTHOR(S)
0.1	Initial draft	01/09/2018	Vladimir Alexiev (ONTOTEXT)
0.7	Complete draft	17/09/2018	Vladimir Alexiev (ONTOTEXT)
0.8	Input from partners	21/09/2018	Franco Maria Nardini (CNR), Raffaele Perego (CNR), Nicola Tonellotto (CNR), Pythagoras Karampiperis (Agroknow), Antonis Koukourikos (Agroknow)
0.9	Internal Review	28/09/2018	Franco Maria Nardini (CNR)
1.0	Final edits after internal review	04/10/2018	Vladimir Alexiev (ONTOTEXT)

PARTICIPANTS		CONTACT
Agroknow IKE (Agroknow, Greece)		Pythagoras Karampiperis Email: pythk@agroknow.com
Ontotext AD (ONTOTEXT, Bulgaria)		Vladimir Alexiev Email: vladimir.alexiev@ontotext.com
Consiglio Nazionale Delle Richerche (CNR, Italy)		Raffaele Perego Email: raffaele.perego@isti.cnr.it
Katholieke Universiteit Leuven (KULeuven, Belgium)		Katrien Verbert Email: katrien.verbert@cs.kuleuven.be
Geocledian GmbH (GEOCLEDIAN Germany)		Stefan Scherer Email: stefan.scherer@geocledian.com
Institut National de la Recherché Agronomique (INRA, France)		Pascal Neveu Email: pascal.neveu@inra.fr
Agricultural University of Athens (AUA, Greece)		Katerina Biniari Email: kbiniari@aua.gr
Abaco SpA (ABACO, Italy)		Simone Parisi Email: s.parisi@abacogroup.eu
APIGAIA (APIGEA, Greece)		Eleni Foufa Email: Foufa-e@apigea.com

ACRONYMS LIST

AEO	Agricultural Experiments Ontology
AFEO	Agri-Food Experiment Ontology
AGRO	AgroKnow
AGRO	Agronomy Ontology
AgroBio	Agronomy and Biology data
AT	Agricultural Technology Ontology
AUA	AGRICULTURAL UNIVERSITY OF ATHENS
BCO	Biological Collection Ontology
BFO	Basic Formal Ontology
ChEBI	Chemical Entities of Biological Interest
CLO	Cell Line Ontology
CO	CropOntology: a group (set) of ontologies for specific crops
CO_320	CropOntology: Rice
CO_321	CropOntology: Wheat
CO_322	CropOntology: Maize
CO_356	CropOntology: Vitis (grapes/viticulture)
CO_357	CropOntology: Woody Plants
CSV	Comma-Separated Values
CUBE	W3C ontology for representing multidimensional data cubes
DC	Dublin Core (elements)
DCT	Dublin Core Terms
DOID	Human Disease Ontology
EBI	European Bioinformatics Institute
EC	Electrical conductivity
ECA	Eddy Current Array
EDAC	Earth Data Analysis Center (data produced by Earth, Life and Semantic Web project)
EFO	Experimental Factor Ontology
EM-38	A handheld Geonics electromagnetic soil conductivity meter
EMI	Electromagnetic Induction: used in soil conductivity sensors (see also ECA)
ENVO	Environment Ontology
EO	Environment Ontology
eyeball	A Jena tool for RDF semantic validation (e.g. that no unknown terms are used)
FOODON	Food Ontology
GeoSPARQL	Geographic extensions to SPARQL. Defines representing features, geometries (e.g. asWKT) and spatial relation predicates (e.g. sfContains)
GIS	Geographic Information System
GODAN	Global Open Data for Agriculture and Nutrition
GPS	Global Positioning System

GraphDB	Semantic repository (database) by ONTO
grlc	Git Repository Linked data API Constructor
HDOP	Horizontal Dilution of Precision of a GPS reading
HTML	W3C HyperText Markup Language
IAO	Information Artifact Ontology
INRA	Institut national de la recherche agronomique
LAI	Leaf Area Index
LIRMM	Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier
LOV	Linked Open Vocabularies, a site for discovering ontologies
MMO	Measurement Methods Ontology
NASA	National Aeronautics and Space Administration
NCBITaxon	NCBI Taxonomy
NDRE	Normalized Difference Red Edge
NDVI	Normalized Difference Vegetation Index
NIR	Near-infrared spectral region
NIRi	Incident radiation of the Near-InfraRed spectrum
NIRr	Reflected radiation of the Near-InfraRed spectrum
OBO	Open Biological and Biomedical Ontology
OEPO	Ontology for Experimental Phenotypic Objects
OFPE	Ontology for Food Processing Experiment
OLS	Ontology Lookup Service
OWL	W3C Web Ontology Language, a more complex language for describing ontologies
OxO	Ontology Xref (Cross-Reference) Service
PATO	Phenotypic Quality Ontology
PCO	Population and Community Ontology
PECO	Plant and Environmental Conditions Ontology
PO	Plant Ontology
QB	See CUBE
QUDT	NASA Quantities, Units, Dimensions, and Types Ontology
R	Red spectral region
RDA	Research Data Alliance
RDBMS	Relational Database Management System
RDF	W3C Resource Description Framework, the semantic web data model
RDF Shapes	A way to describe semantic data Application Profiles. Two approaches are SHACL and ShEx
rdfpuml	ONTOTEXT tool for translating RDF to PlantUML, a textual notation for generating UML diagrams
RDFS	W3C RDF Schema, a simple language for describing ontologies
RE	Red-Edge spectral region (spectrum centred around 715 nm)
REDi	Incident radiation of the red spectrum
REDr	Reflected radiation of the red spectrum
REST	Representational State transfer

RIOT	RDF Input/Output Tool, part of Apache Jena. Includes RDF syntax validation
RO	Relations Ontology
SDGIO	SDG-Interface Ontology
SHACL	Shapes Constraint Language, a W3C Recommendation
ShEx	Shape Expressions, a W3C community specification
SKOS	Simple Knowledge Organization System, an ontology for describing thesauri
SPARQL	SPARQL Protocol and RDF Query Language
TO	Trait Ontology
TSV	Tab-Separated Values
Turtle	Terse RDF Triple Language
UML	Unified Modeling Language
UO	Units Ontology
URL	Uniform Resource Locator
VANN	Vocabulary for annotating vocabulary descriptions
W3C	World Wide Web Consortium
WKT	Well-Known Text, a format for describing feature geometries
WP	Work package
XML	W3C eXtensible Markup Language
XO	Experimental condition ontology
XSD	XML Schema Datatypes

EXECUTIVE SUMMARY

WP3 Data & Semantics Layer is a core WP of the project. If we have no data, we cannot achieve almost any of the project objectives. Within this WP3, task T3.1 Data Modelling over Big Data Infrastructures has these objectives:

- Explores partner data
- Defines competence questions that the data should be able to answer
- Studies relevant AgroBio ontologies
- defines semantic modelling principles and specific models
- Studies user (researcher) requirements for discovering ontologies, mapping data, aligning data, etc.
- Implements or adopts tools for these requirements

The document has the following structure:

- Chapter 1 Introduction describes fundamental AgroBio data (observations and measurements), outlines the ontological representation of measurements, mentions possible alternatives (e.g. following existing AgroBio patterns vs using the W3C CUBE ontology), describes the steps of semantic data integration, and provides links to consortium resources related to the task.
- Chapter 2 Relevant AgroBio Ontologies outlines the vast number of potentially relevant ontologies and the terms included in them. We provide some metrics (number of terms) and surveys various Ontology Portals and Tools that are available for browsing, finding and using ontologies; and that can also serve as inspiration for developing requirements for tools to be developed/adopted by the project.
- Chapter 3 Improving AgroBio Ontologies describes a variety of problems that we have found in AgroBio ontologies, and the initial steps we have taken to engage with the AgroBio communities to improve the quality of these ontologies. We also show a case of searching for a specific term (NDVI) required by specific partner data in a couple of ontology portals.
- Chapter 4 Specific Project Data discusses specific consortium data (including problems of draft semantic data that will be corrected), data processing requirements and data access requirements.
- Chapter 5 Conclusions provides conclusions, next steps and a bibliography.

Deliverable D3.1 Data Modelling and Linking Components will have 3 iterations at M9, M21, M30. In this first iteration (M9), we describe the first steps taken for the realization of task T3.1. These initial steps were taken to clarify the scope and essential ingredients of the task.

Since the project is early in its life cycle, we do not yet have finalized requirements for the tools to be developed by Task 3.1. Section 1.4 outlines the steps that we intend to follow, and the approximate point that we have reached within these steps.

TABLE OF CONTENTS

LIST OF FIGURES

1 INTRODUCTION

Deliverable D3.1 is defined as "A tool for creating, maintaining and linking semantic data, customized to serve the needs of the relevant grapevine-powered industries".

This deliverable is part of task T3.1, which is described as:

- Work on the task will initially focus on the provision of a basic integrated model for grapevine-powered industries, facilitating interoperability between the data assets of the different industries and incorporating open data from third-party entities that pertain to use cases specified in T2.1.
- Consequently, the BigDataGrapes model will be published as an ontology, and linked with external conceptualizations via a semi-automatic process. The scalable ontology alignment systems envisioned in the project will be implemented and applied for linking the model with significant specifications, either general purpose or domain-specific.
- Furthermore, the task will produce the necessary tools and components for carrying out the aforementioned processes, i.e. an environment for building, reusing and linking disparate conceptualizations.

D3.1 will have 3 iterations at M9, M21, M30. During this first period we worked on the first bullet of the task, namely:

- conducted a series of project meetings (2 face-to-face and 6 online)
- explored relevant AgroBio ontologies
- studied the datasets provided by the partners
- discussed different conceptualizations and semantic data models for representing this data
- various approaches for structuring semantic data models, including ontologies and application profiles. Approaches for describing data models, including UML diagrams and RDF Shapes
- discussing Competence Questions that the semantically integrated data should be able to answer.
- discussing user (researcher) needs regarding the discovery and selection of ontologies, searching and selecting classes and properties, mapping tabular data to RDF, etc.
- found a number of problems in the reference AgroBio ontologies and engaging with the relevant user communities for fixing those problems

1.1 FUNDAMENTAL AGROBIO DATA: MEASUREMENTS

The basic data that needs to be represented by the project is AgroBio **measurements/observations**: the measurement of some traits of some objects (e.g. soil or a particular crop) using a certain method, technique, equipment, units of measure, time, place, etc. This sounds simple, but it involves a number of data items to give the observations context and meaning.

We can illustrate it with an example regarding measuring a basic variable: plant height.

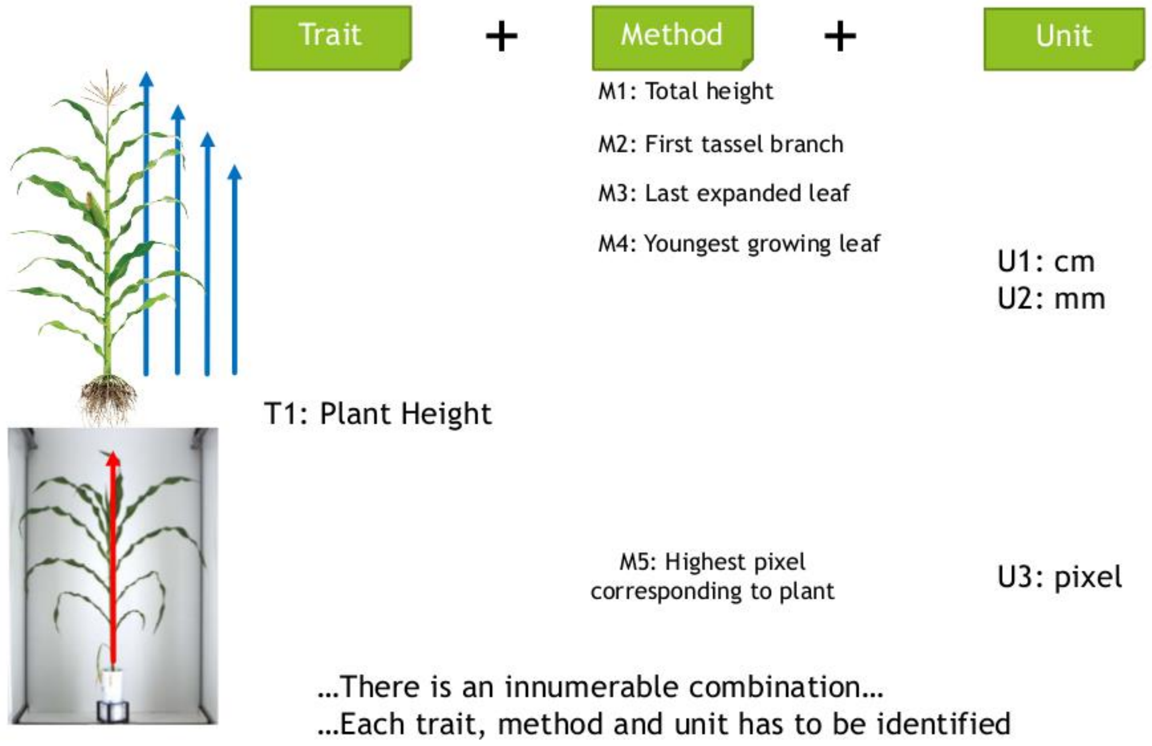

Figure 1 Basic Measurement: Plant Height

A measurement involves the following items:

- **Entity**: thing being measured or observed, such as the soil, weather (temperature, precipitation), the soil, a particular crop or plant variety, harvest parameters, etc.
- **Quality**: what is being measured
- **Trait** = entity + quality: what quality of which entity
- **Method**: what exactly are we measuring (e.g. height to youngest growing leaf or total plant height) and how (instrument, technique, etc).
- **Unit** of measure: may include fundamental units (e.g. Meter, Second), derived units (e.g. m/s) or a variety of countable units (e.g. pixels, count, etc).
- **Variable** = trait + method + unit: provides the detailed meaning of the measurement.
- **Context**: circumstances of the observation, e.g. GPS location, estate/plot/subplot, depth of measurement (for soil), datetime, etc. May also include qualifiers, e.g. instrument, which satellite provided GPS location, precision, instrument status at the time the reading was taken, whether there's a metal pole at the location (which makes a conductivity measurement invalid, who took the reading, etc.
- **Value**: the number that was measured/observed
- **Observation** = variable + context + value: all details about a single observation point.

Please note that it is a common practice to measure several variables of the same entity at once (in the same context). Combination instruments make this possible, and it saves time and effort. This leads to the need to share entity and context between observations, which affords the following efficiencies:

- Easier correlation of related observations
- More economical data representation

1.2 ONTOLOGICAL REPRESENTATION OF MEASUREMENTS

There are various different ways to represent AgroBio measurements using the RDF semantic data model, two of which are:

- Using some of the established AgroBio ontologies. The next chapter introduces such ontologies, but we give below a motivating example of measuring plant height.
- Using the W3C CUBE ontology for representing multidimensional observations, which is described in the next subsection.

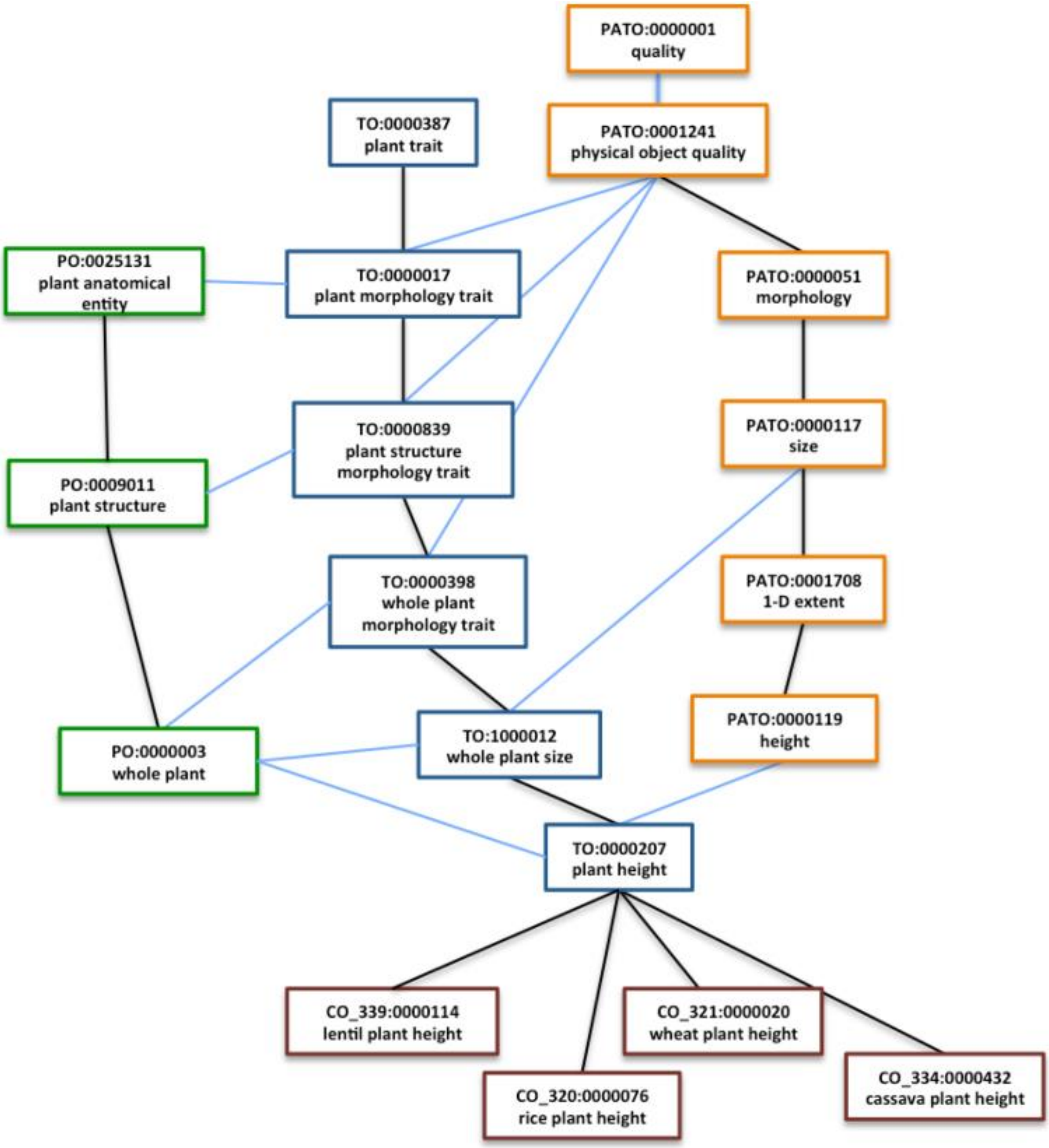

Figure 2 Semantic Classes for Representing Plant Height

- The Phenotypic Quality Ontology (PATO, orange chain) is used to classify the trait (considered as a physical object quality) in a subsumption hierarchy.

- The Plant Ontology (PO, green chain) is used to describe plant anatomical parts, i.e. sub-entities that can be measured
- The Trait Ontology (TO, blue chain) is used to describe a particular plant morphology, i.e. tie the trait to an anatomical part
- The Crop Ontologies (CO_nnn, brown subclasses) specialize the trait to particular crops or varietals

We identify a problem with tying up a trait that is quite universal (height) to such a specific degree. A height is a height, no matter whether you measure lentils, rice, wheat, any plant, or a skyscraper. It's true that measurement methods often vary per entity, i.e. are applicable only to certain kinds of entities. But that restricted applicability does not mean that every variable should be replicated to every crop that it applies to, which leads to a combinatorial explosion.

We locate this problem (improper level of abstraction) many times in the Crop Ontologies, for example:

- Normalized difference vegetation index (NDVI) is defined in CO_322 Maize, so we can't use it for Grapes. NDVI is not defined in CO_356 Vitis. But rather than replicating NDVI in Vitis, its proper place is in the general Crop Ontology (CO), not a sub-ontology of CO.
- The "grams" unit of mass is bound to some Woody Plant trait, so we can't use it for Grapes.

We believe that by "regrouping the factors" in the equations outlined in section 1.1, we can avoid such combinatorial explosions:

- **Current**: Trait = entity + quality; Variable = trait + method + unit; Observation = variable + context + value
- **Future**: Variable = quality + method + unit; Observation = entity + variable + context + value. Quality defines which entities it is applicable to but is not subjugated to Entity.

1.2.1 Creating New Ontology Terms

We will often need to define new terms. Some example:

- There is nothing about "the number of grapes" in Vitis. We could create this trait using "grape" in Agricultural Experiments Ontology (AEO) and "amount" in Phenotypic Quality Ontology (PATO).
- CV1m: "soil conductivity at depth 1 meter in millisiemens per metre (mS/m)" encapsulates 4 factors: entity: soil; variable: electrical conductivity; context: depth=1m; unit: mS/m. There's nothing about soil conductivity in CO or Vitis. The closest we can find is ENVO:09200016 conductivity of soil. We could use that, and then construct extra terms to specify the unit (mS/m) and context (1m vs 0.5m depth)
- The closest we can I find to specific-spectrum measurements (Near-infrared, Red, Red-Edge) is FIX:0000641, but that only has "far-, mid- and near-infrared spectroscopy". For some AUA data (see sec 4.2.2) we need to express more specific spectrum measurements.
- We can find NDVI in CO_322 Maize, but not in CO_356 Vitis. Should we create another term "NDVI for grapes", thus perpetuating the increase of number of terms? We believe that CO should define NDVI in a crop-independent manner, then we can just use that rather than making a number of crop-dependent terms.

However, the traditional approach of creating new terms for every combination will possibly lead to a combinatorial explosion in the number of terms. If we vary any one of these factors, we will need another term.

1.3 W3C CUBE ONTOLOGY

The W3C CUBE ontology (QB) captures multidimensional observations (data cubes) using the following terminology (in bold). We roughly map these QB terms to the data items discussed in previous sections. The first 3 are called "components".

- **Dimension:** entity, quality, method
- **Attribute:** unit, context
- **Measure:** value
- **Observation** = Dimensions + Attributes + Measures

QB defines what components are expected in a specific qb:DataSet by using a qb:DataStructureDefinition. QB provides some flexibility that affords data efficiencies, and avoiding combinatorial explosion:

- QB allows using several dimensions per observation, without tying them up together. E.g. you can use 3 dimensions entity="plant", quality="height", method="whole height"
- We could also use several measures per observation (e.g. as taken by a combination instrument), although this is less commonly used.
- One could split a dataset into Slices (or other kinds of ObservationGroups) by fixing some of the dimensions, so one doesn't need to repeat them for every observation.

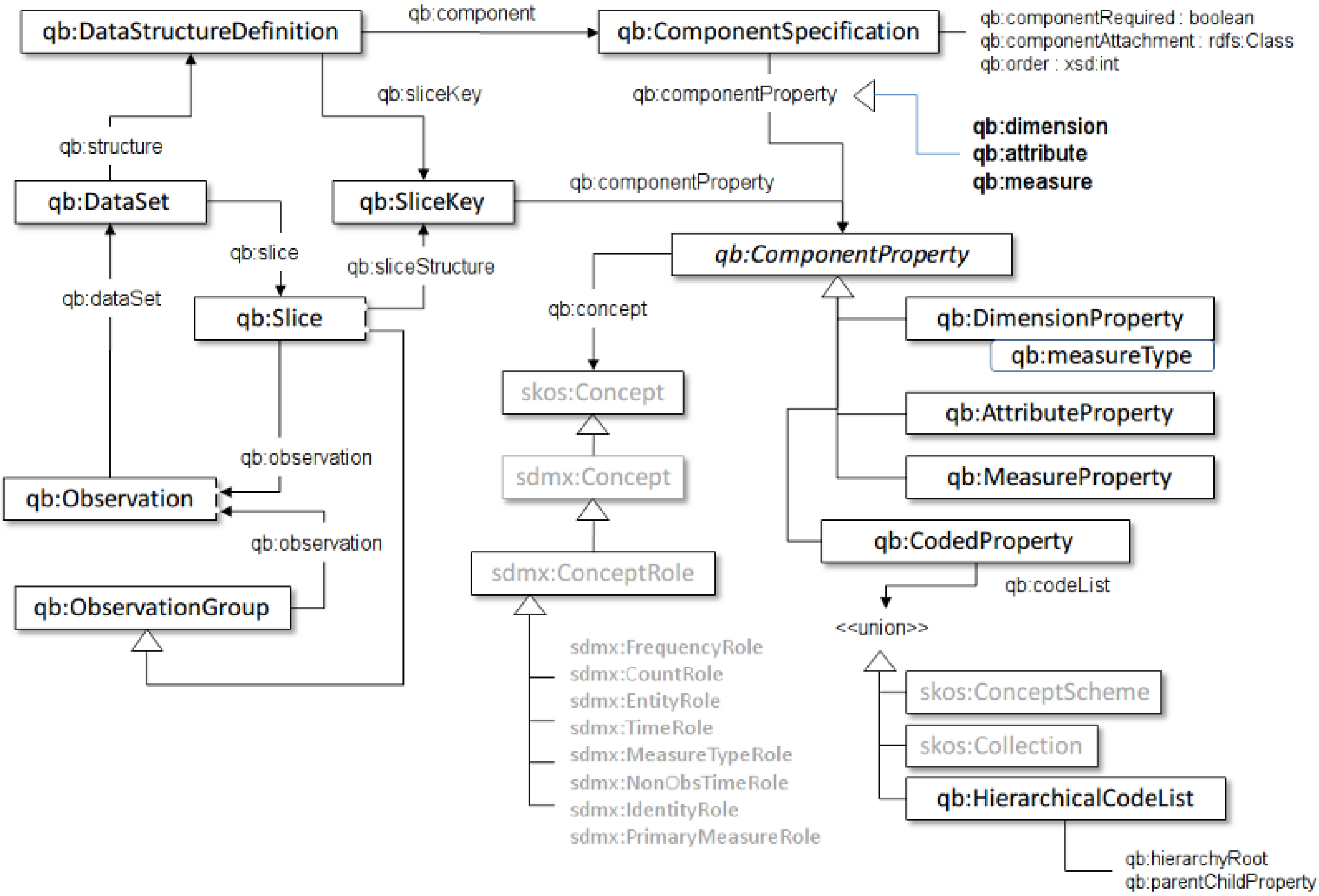

Figure 3 W3C CUBE Ontology

1.4 SEMANTIC DATA INTEGRATION

Semantic Data Integration has proven itself in the last 10 years as one of the best ways to integrate diverse data across institutions and enterprises, and to leverage datasets available in the LOD cloud. Life Science and Biology researchers were one of the early adopters of semantic web techniques, and by now they have found a wide following also in the Agricultural community, who in many cases leverages ontologies developed in the Bio community.

Semantic Data Integration is a holistic activity that aims to harmonize data from different providers, convert it to a semantic form, match (coreference) instances about the same entity coming from different datasets, and create an integrated Knowledge Graph of data in a domain. It involves the following steps, which have informed and will continue to inform WP3 activities:

- Get sample tabular data from partners
- Get sample RDF data from partners
- Analyse the data
- Define competence questions and other data requirements
- Research ontologies sent by partners and other related ontologies
- Report ontology and instance data errors to partners and the AgroBio ontology
- Ontology engineering: selection, combination and extension of ontologies

The consortium's progress to date is somewhere at this point.

- Discuss how to represent various data aspects with partners: estates/plots, measurements/observations, equipment, experiments, etc
- Create a semantic model with rdfpuml and text narrative (see the euBusinessGraph Semantic Model as an example)
- Get the model approved by all partners
- Create application profiles and/or RDF shapes (SHACL and/or ShEx) for validation of semantic data for conformance to the model
- Define URL design and policies
- Semantic conversion using appropriate tools depending on source (CSV/TSV tabular, RDBMS, XML)
- Semantic alignment and instance matching
- Data validation and data quality management/measurement
- Implement proper semantic publishing and content negotiation
- Design and implement data update flows
- Create sample queries
- Deploy predefined queries as REST services

1.5 RESOURCES

We have created a public GitHub repository https://github.com/BigDataGrapes-EU/ontology for WP3 work. For now, it has the following folders:

- data: semantic data (for now mostly 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. In particular, see README: Github preview and Rendered HTML version

We also have a Google Drive folder that is available only to consortium members for carrying out intermediate communication before reaching to publishable results.

2 RELEVANT AGROBIO ONTOLOGIES

A very large number of ontologies have been developed in the AgroBio domain, starting with biological and life science related ontologies, and continuing into the agronomy and crop science domains. We have studied these ontologies in order to make informed choices in our semantic modelling, and to reuse existing ontologies as much as possible.

The initial set comprised 16 relevant ontologies, plus 21 on specific crops (Vitis on grapes, and 20 more that we could use as examples). Out of these candidates, we downloaded 17 ontologies, converted from RDF and OWL to Turtle (which is easier to read), added prefixes where missing, assessed ontology coverage and quality, and wrote up various problems that we encountered (see next chapter).

2.1 RESEARCHED ONTOLOGIES

The ontologies we researched include:

- 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_356: Vitis (viticulture)
- CO_320: Rice
- CO_322: Maize
- CO_357: Woody Plant
- CO_UO: Units Ontology
- EO (ENVO): Environment Ontology
- FOODON: Food 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 Environmental Conditions Ontology
- PO: Plant Ontology
- RO: Relations Ontology
- SDGIO: SDG-Interface Ontology
- TO: Trait Ontology
- UO: Unit Ontology
- XO: Experimental condition ontology

2.2 ONTOLOGY METRICS

The set of ontologies researched is very large, i.e. includes a large number of ontological terms. ("Individuals" refers to resources that are neither classes nor properties, e.g. lookup values.)

Table 1 AgroBio Ontologies and Number of Terms

Ontology	Classes	Properties	Individuals
AEO	56	36	30
AFEO	68	8	0
AGRO	1685	709	284
BCO	157	279	28
BFO	35	20	0
CHEBI	128900	45	0
CO	0	0	0
CO_356 Vitis	814	10	0
ENVO	8510	241	21
FOODON	27050	130	359
IAO	219	111	23
NCBITaxon	1692930	27	0
OEPO	110	60	0
OFPE	71	149	58
PATO	2713	61	0
PCO	242	636	20
PECO	2974	137	0
PO	2000	63	0
RO	80	650	0
SDGIO	514	171	702
TO	5041	201	0
UO	420	35	0
XCO	535	28	0
Total	**1875124**	**3807**	**1525**

These counts are not completely accurate for several reasons:

- Many ontologies describe "foreign" terms (from external namespaces). Not only this constitutes "namespace hijacking", but it also likely inflates the metrics.
  - E.g. AFEO and AEO include fairly complete copies of SKOS and DCT.
  - E.g. FOODON includes 97 foreign terms from CEPH, CHEBI, ENVO, GAZ, GENEPIO, IAO, NCBITaxon, OBI, PATO, PO, RO, UBERON.
- Different numbers are reported for some of the ontologies, e.g.

Table 2 Differences in Ontology Term Counts

Ontology	Classes	Properties	Individuals	Source
AEO	56	36	30	http://agroportal.lirmm.fr/ontologies/AEO
AEO	56	36	0	http://vest.agrisemantics.org/content/agricultural-experiments-ontology
AEO	250	26	0	http://www.ontobee.org/ontostat
TO	5041	201		http://www.ontobee.org/ontostat
TO	4927			https://www.ebi.ac.uk/ols/ontologies/to

In any case, it is evident that the complexity of these ontologies is very high. This requires the use of portals and tools to investigate relevant ontologies and the find relevant terms (see next subsection). We researched several ontology portals that hold a total of:

- 200 ontologies,
- 5M classes,
- 16k properties,
- 476k individuals

2.3 ONTOLOGY NOTES

This section includes brief notes about some specific ontologies.

2.3.1 Class Information Template

- rdfs:label: name
- obo:IAO_0000115: definition
- rdfs:comment: scope note
- rdfs:subClassOf: subclasses and restrictions
- owl:equivalentClass: restrictions (eg owl:intersectionOf)
- oboInOwl:created_by
- oboInOwl:creation_date
- oboInOwl:hasDbXref: sources of information (commit, author or publication), e.g. "NIG:Yukiko_Yamazaki", "FNA:00e30ce4-70bc-489c-86df-73030c9ece1e", "PO_GIT:658", "PO_REF:00002", "POC:curators", "ISBN:9780023681905", "PMID:18978364", "GO:0022611"
- oboInOwl:hasExactSynonym
- oboInOwl:hasOBONamespace
- oboInOwl:id

2.3.2 Deprecated Classes

Deprecated classes are expressed as follows:

```
obo:PO_0006441 a      owl:Class ;
    obo:IAO_0000231  obo:IAO_0000227 ;
    obo:IAO_0100001  obo:PO_0009029 ;
    owl:deprecated   true .
```

2.4 ONTOLOGY PORTALS AND TOOLS

As part of reviewing related work, we have researched several ontology portals and related tools that can be useful to the project as follows:

- As tools for finding and researching ontologies
- As tools for finding relevant terms within ontologies
- As useful examples to inspire the creation of BigDataGrapes tool requirements

This research could be elaborated by going through additional lists of tools. For example, 25 tools, datasets and ontologies were presented during the PhenoHarmonIS 2016 workshop.

2.4.1 OBO Foundry

OBO Foundry (Open Biological and Biomedical Ontology Foundry) is the largest portal listing relevant ontologies, with resource links (license, detailed info, project home, issues, developer contacts, download, Ontobee browser).

Figure 4 OBO Foundry Resource Links

2.4.2 CropOntology Portal

The Crop Ontology Curation Tool is a collaborative ontology development portal that allows visualization and submission of ontologies. It includes a hierarchical tree browser, per-term metadata, and a graph visualization of the neighbourhood of a term, though the latter is relatively unclear and non-intuitive.

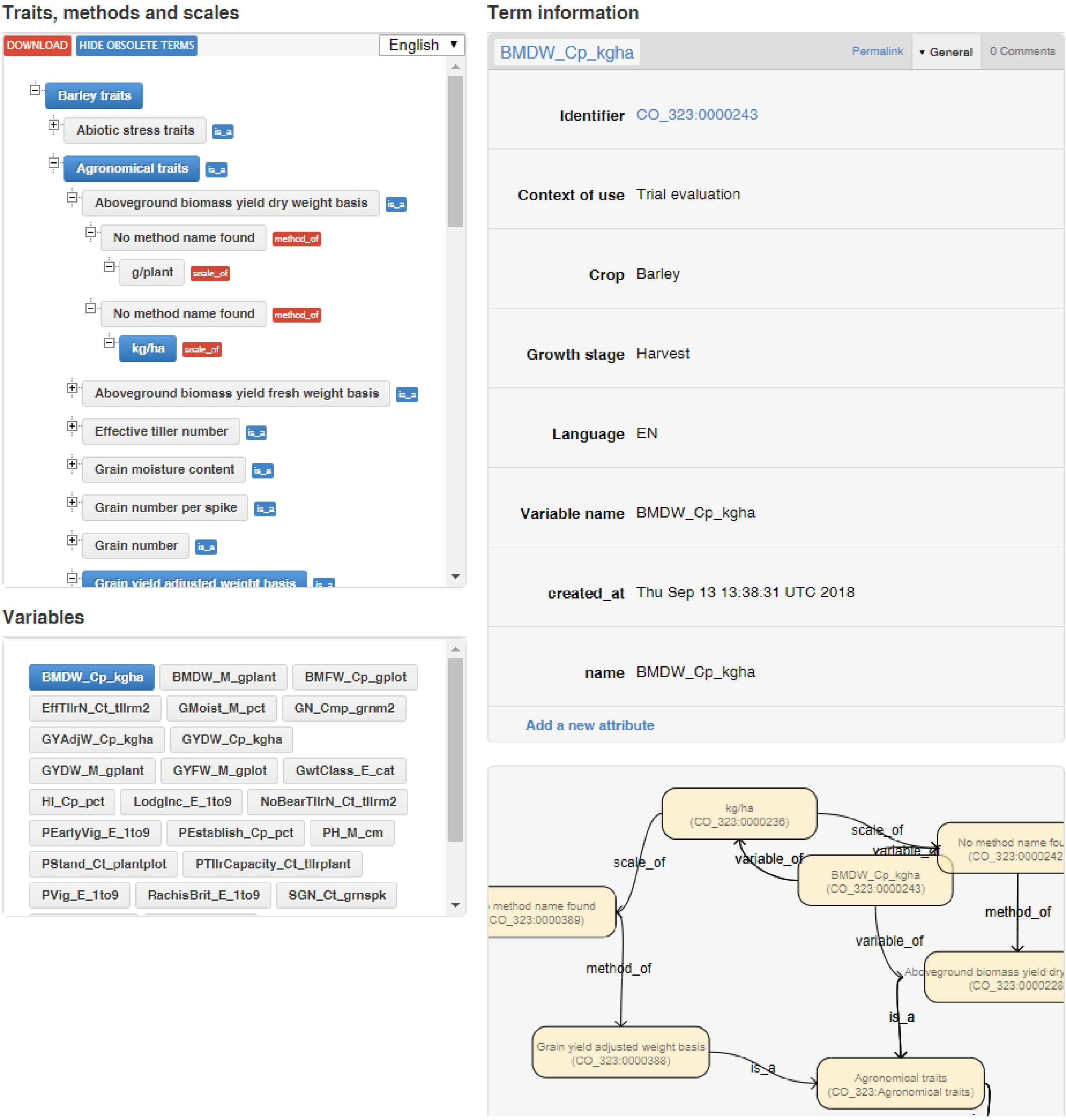

Figure 5 Crop Ontology Showing a CO_323 Barley Term

2.4.3 CropOntology Annotation Tool

The CropOntology Annotation Tool permits the annotation of tabular data with terms from AgroBio ontologies. After loading some tabular data, the headers are matched against a number of ontologies and all matches are suggested as possible selections. One can also limit by kind of crop. Here we have loaded some AUA table grapes data and matched one of the fields (NDVI). The tool uses only literal matches and very often there is a lot of ambiguity in selecting correct fields.

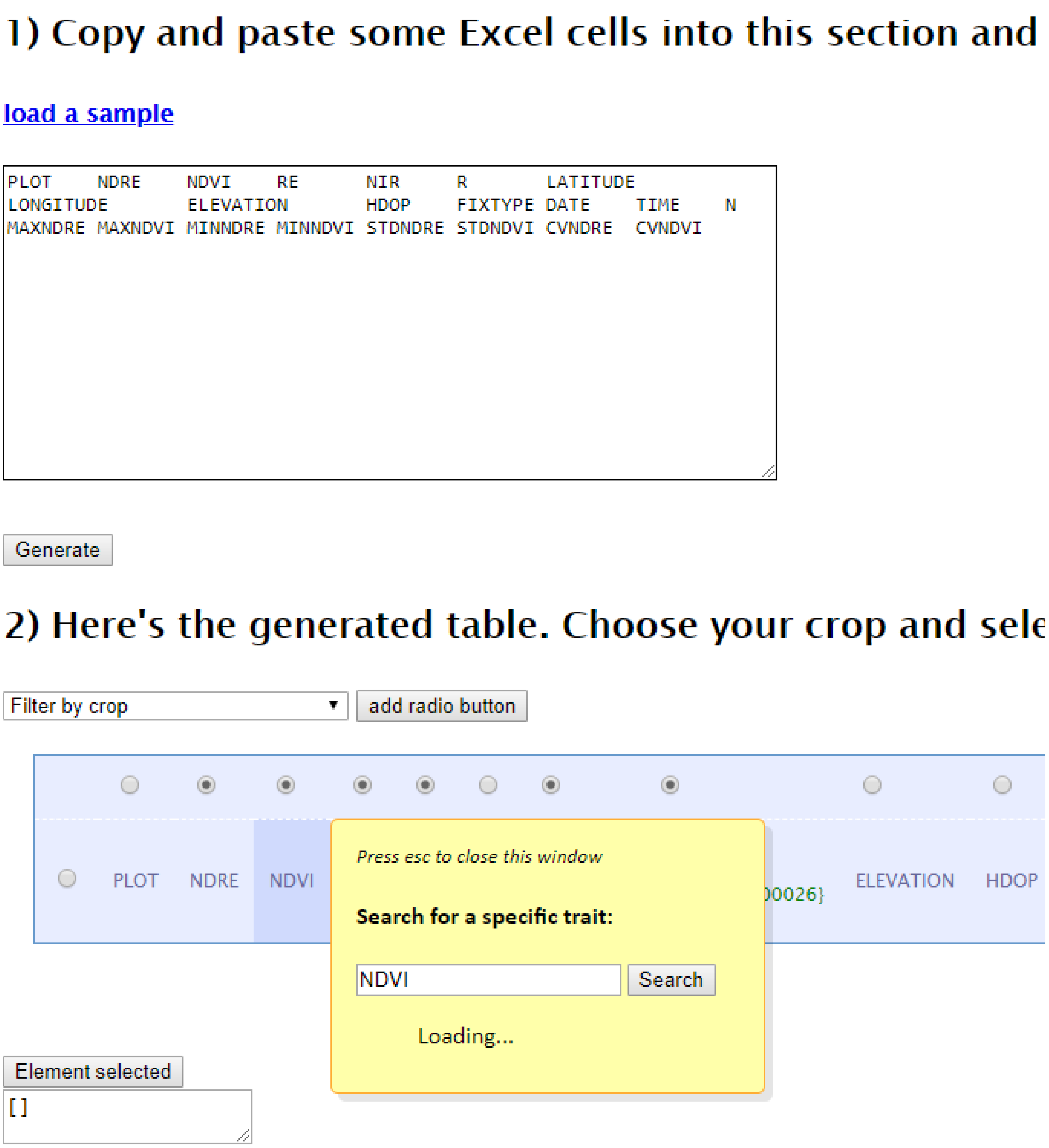

Figure 6 CropOntology Annotation Tool

2.4.4 Planteome

Planteome Browser (AMIGO) allows detailed browsing for terms in PO, TO, EO, PSO, GO, PATO, CHEBI.

- It includes a tree browser (drill-down) of 2M bio-entities that is a great aid in understanding the hierarchical structure of an ontology.

Figure 7 Planteome Browser Tree Drill-Down

- Includes tree view (hierarchical position) and graph visualization of the selected term.

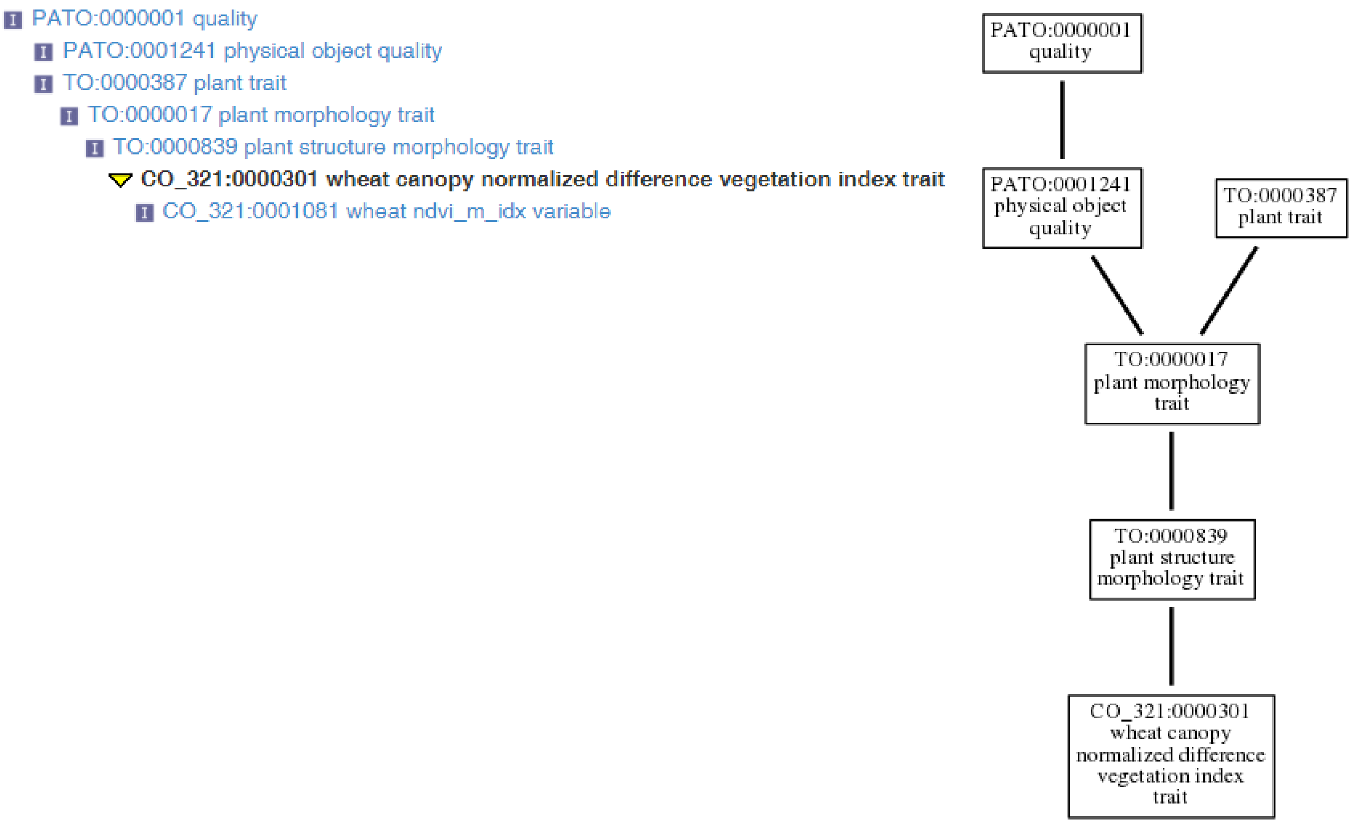

Figure 8 Hierarchical Tree and Graph of Term CO_321:0000301 wheat canopy NDVI trait

2.4.5 EBI OLS

The EBI Ontology Lookup Service aggregates 200 ontologies, 5M classes, 16k props, 476k individuals (according to tweets @EBIOLS).

- It includes a tree browser for classes and properties, and shows graph visualizations
- It's one of the few portals that includes the Crop Ontologies
- Includes an extremely useful faceted search

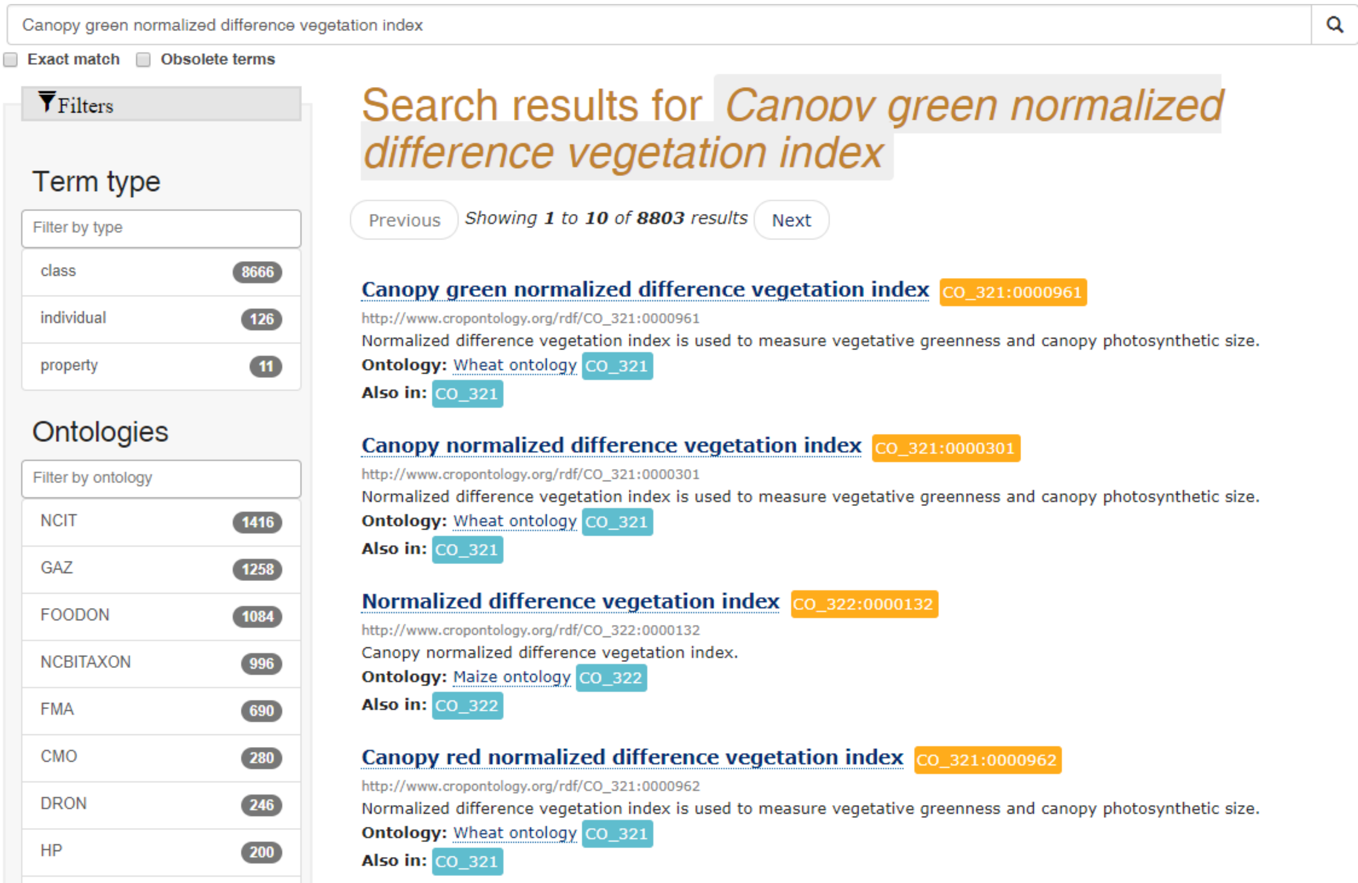

Figure 9 EBI OLS Faceted Search

2.4.6 EBI OxO

The EBI Ontology Xref Service (OxO) allows exploration of ontology mappings between all ontologies included in OLS, and some UMLS ontologies. It starts with an overview of available mappings, after which you can see statistics of mappings from/to a selected ontology.

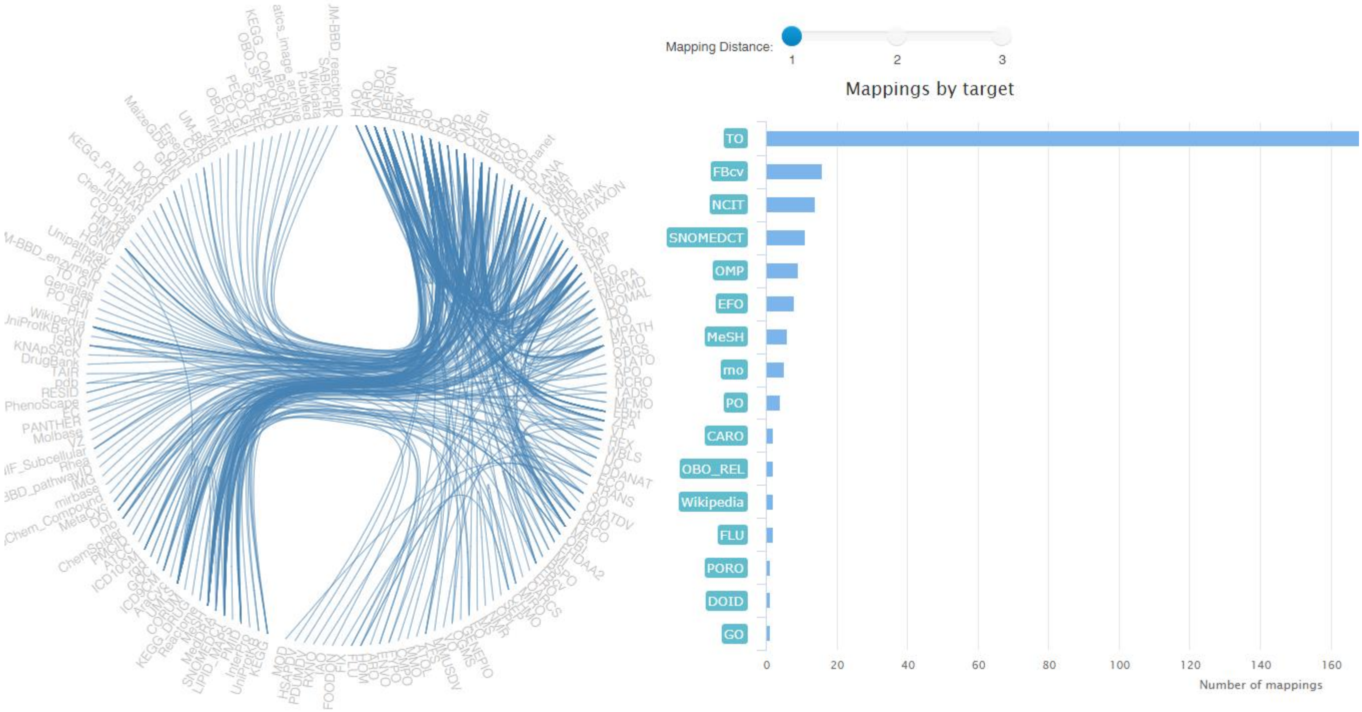

Figure 10 EBI OxO Home Page and Selected Target Ontology (PATO)

EBI OxO also allows the exploration of the mappings of a list of terms, including inference of indirect paths and evidence (in which ontology the mappings were found).

Showing 1 to 50 of 137 entries   Search:

Input	Mapped id	Id source	Evidence	Distance
DOID:162 (cancer)	MedDRA:10028997 (Neoplasm malignant)	MedDRA	7	2
DOID:162 (cancer)	SNOMEDCT:269626006 (Ca - unspecified site NOS)	SNOMEDCT	6	1
DOID:162 (cancer)	SNOMEDCT:269513004 (Malignant neoplasms (& carcinoma))	SNOMEDCT	6	1
DOID:162 (cancer)	UMLS:C2242772 (Malignant neoplasm NOS (& sarcoma NOS))	UMLS	7	2
DOID:162 (cancer)	MONDO:0005070 (neoplasm (disease))	MONDO	7	2
DOID:162 (cancer)	MONDO:0004992 (cancer)	MONDO	1	1
DOID:162 (cancer)	SNOMEDCT:190150006 ([X]Malignant neoplasm without specification of site)	SNOMEDCT	6	1

Figure 11 EBI OxO Mappings for Selected Terms

2.4.7 AgriSemantics VEST

The AgriSemantics Map of Data Standards (also known as VEST) was developed by the Global Open Data for Agriculture and Nutrition (GODAN) action.

- It includes 398 ontologies with detailed info (e.g. see VEST record on AEO)
  - 215 relate to Food and Agriculture; 180 are Generic / peripheral
  - 76 come from the LIRMM AgroPortal, 328 from GODAN's own VEST Registry

- It includes a search and a faceted Advanced Browse by domain, e.g. there are 55 ontologies on Plant Science and Plant Products. The facets include domain, standard type (e.g. ontology vs taxonomy), format, and an Assessment part on Content (e.g. whether the standard is complete, authoritative), Adoption (e.g. whether it's used in software and how widely), Usability and Openness

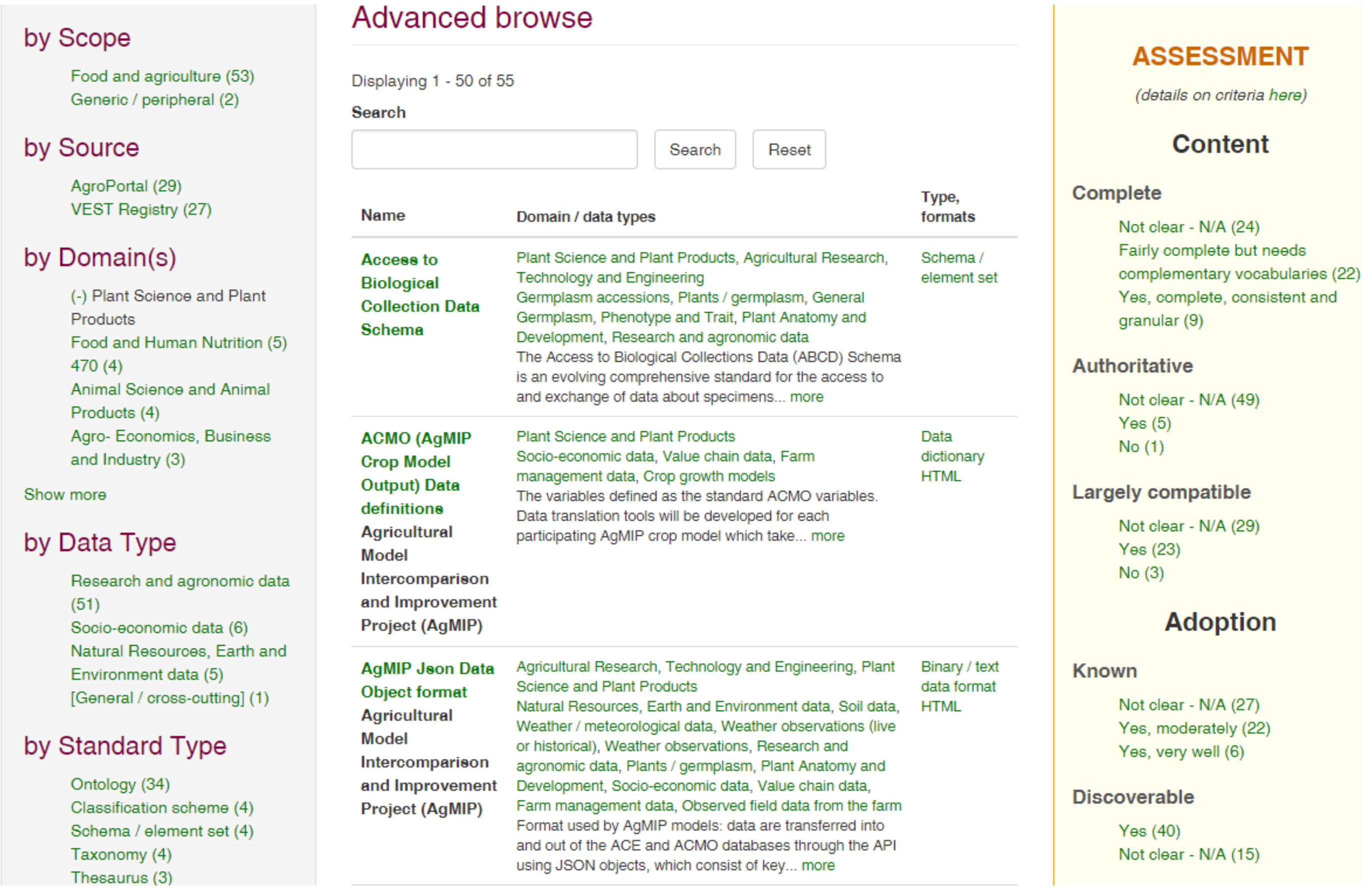

Figure 12 GODAN VEST Advanced (Faceted) Browse

2.4.8 AgroPortal

The LIRMM AgroPortal includes information about 102 Ontologies, 1,734,302 Classes and 1,970,287 Individuals. It includes tools to create ontology-based annotations for your own text , link your own project that uses ontologies to the description of those ontologies , find and create relations between terms in different ontologies, review and comment on ontologies and their components as you browse them, and submit ontology mappings. Data is available for human browsing, and machine consumption through an API or SPARQL endpoint.

- It covers the following ontology sources (so-called "slices"):
  - Crop Ontology Curation Tool (crop)
  - INRA Linked Open Vocabularies (lovinra)
  - OBO Foundry (obo-foundry)
  - The Agronomic Linked Data (AgroLD) (agrold)
  - Consortium of Agricultural Biological Databases (agbiodata)
  - SemanDiv working group (semandiv)
  - RDA Wheat Data Interoperability working group (wheat)
  - Exclusive AgroPortal ontologies (exclu)

- It includes an informative Landscape page with charts and information about group, data catalog, content category, size, most active contributors, etc.

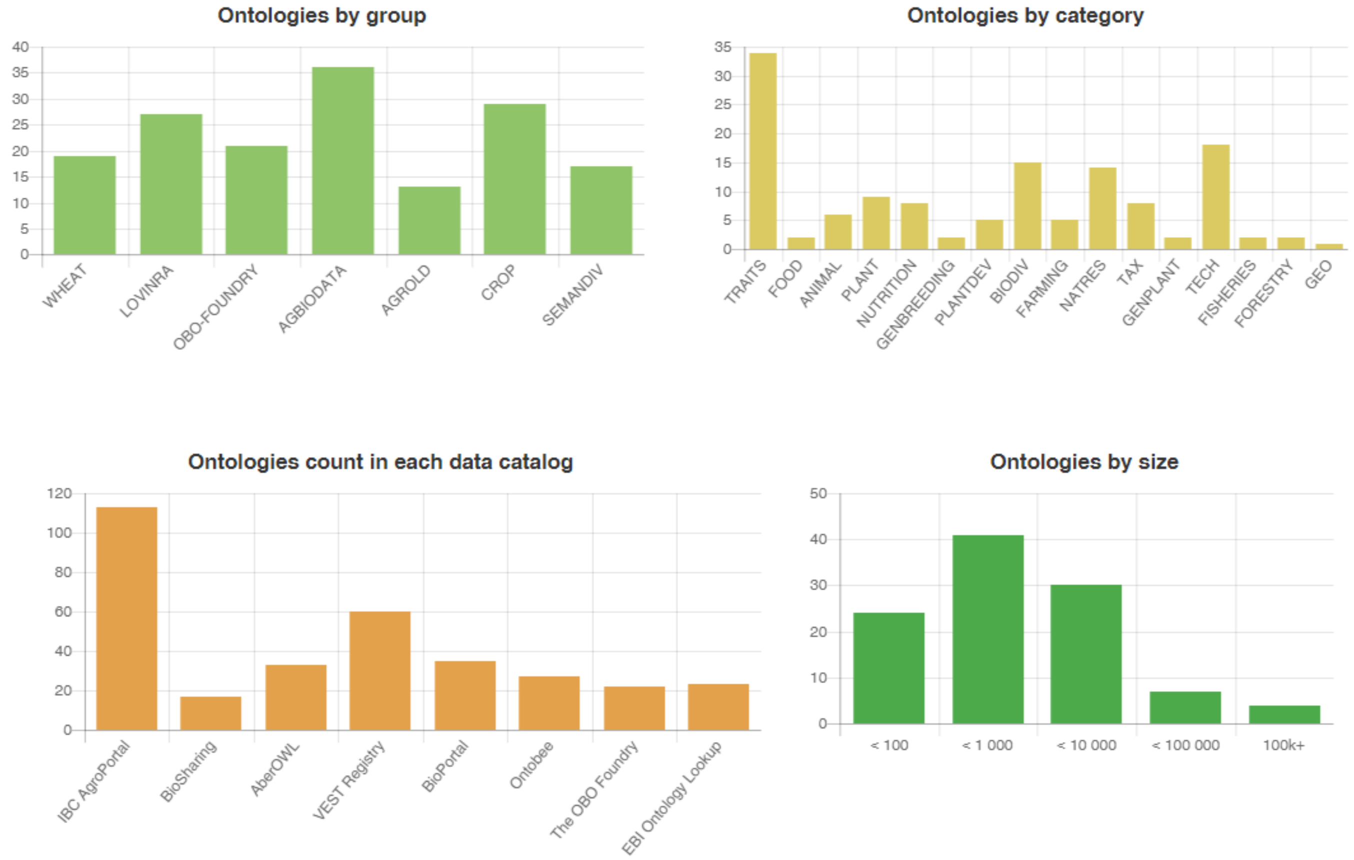

Figure 13 AgroPortal Landscape

2.4.9 OntoBee

OntoBee includes information about 177 ontologies, 4.3M classes, 21k props, 668k individuals (but not the Crop Ontologies).

- It has some of the most detailed statistics per ontology, including imported terms (foreign namespaces) and breakdown of object vs data vs annotation properties.
- Please note that Ontobee treats namespaces as case-insensitive, so these two URLs return the same list: 2200 terms, the union of the two namespaces.
  - http://www.ontobee.org/ontostat/catalog/PO?prefix=PO and
  - http://www.ontobee.org/ontostat/catalog/PO?prefix=po

Ontobee

Home | Intro | Statistics | SPARQL | Ontobeep | Annotator | Tutorial | FAQs | References | Links | Contact | Acknowledge

Statistics of Plant Ontology

Ontology: PO

Index	Ontology Prefix	Class	ObjectProperty	DatatypeProperty	AnnotationProperty	Instance	Total
1	BFO	0	6	0	0	0	6
2	IAO	0	0	0	3	0	3
3	NCBITaxon	11	0	0	0	0	11
4	PO	1,992	0	0	0	0	1,992
5	RO	0	5	0	0	0	5
6	oboInOwl	0	0	0	23	0	23
7	owl	0	0	0	1	0	1
8	po	0	2	0	19	0	21
9	rdf-schema	0	0	0	2	0	2
10	NoPrefix	0	0	0	2	0	2
Total	-	2,003	13	0	50	0	2,066

Figure 14 OntoBee Detailed Statistics about PO

- Includes the most detailed cross-reference of term use across ontologies.

Class Hierarchy

Thing
+ role
  + application
    + pharmaceutical
      - drug
        + cardiovascular drug
        - astringent

Superclasses & Asserted Axioms

- pharmaceutical

This Class is originally defined in

Ontology listed in Ontobee	Ontology OWL file	View class in context	Project home page
Chemical Entities of Biological Interest	chebi.owl	'drug' in chebi.owl	Project home page

Ontologies that use the Class

Ontology listed in Ontobee	Ontology OWL file	View class in context	Project home page
Monarch Disease Ontology	mondo.owl	'drug' in mondo.owl	Project home page
Plant Trait Ontology	to.owl	'drug' in to.owl	Project home page
Human Disease Ontology	doid.owl	'drug' in doid.owl	Project home page
Neuro Behavior Ontology	nbo.owl	'drug' in nbo.owl	Project home page
The Drug-Drug Interactions Ontology	dinto.owl	'drug' in dinto.owl	Project home page
Chemical Entities of Biological Interest	chebi.owl	'drug' in chebi.owl	Project home page
human phenotype ontology	hp.owl	'drug' in hp.owl	Project home page
Gazetteer	gaz.owl	'drug' in gaz.owl	Project home page
Porifera Ontology	poro.owl	'drug' in poro.owl	Project home page
Mammalian phenotype	mp.owl	'drug' in mp.owl	Project home page

Figure 15 OntoBee Cross-Reference of a term (CHEBI:23888 Drug)

- Includes an interesting comparison tool "Ontobeep" that allows to explore term matching and reuse, compares hierarchical ontology structures, and identifies possible redundancy and errors (see tutorial).

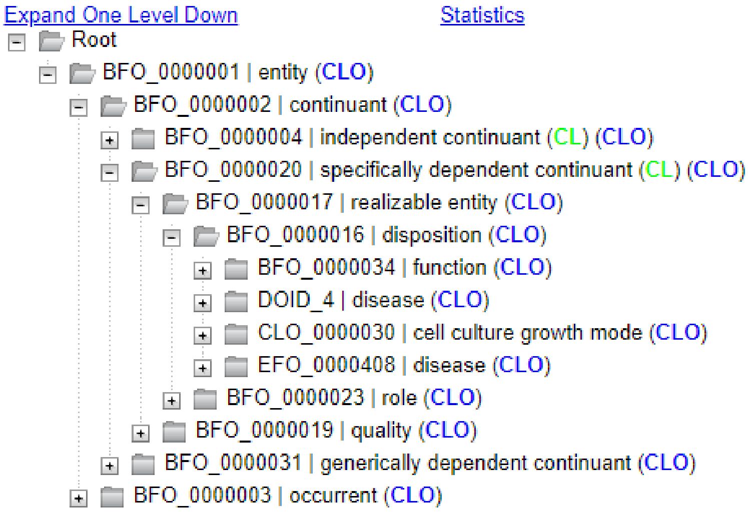

Figure 16 OntoBeep Ontology Comparison

2.4.10 Ontobee Annotator

Ontobee Includes an annotator that can find term occurrences in free text and shows in which ontologies they occur. (The text below is slightly humorous)

Are there any drugs to treat obesity in plants?

Results

Matched Term IRI	Matched Term	Source Ontology	Also available in Ontology
http://purl.obolibrary.org/obo/CHEBI_23888	drugs	CHEBI	AGRO
http://www.ebi.ac.uk/efo/EFO_0001073	obesity	EFO	CLO
http://purl.obolibrary.org/obo/HP_0001513	obesity	HP	

Figure 17 Ontobee Annotator

2.4.11 AberOWL

AberOWL is a simpler ontology browser. For example, the data it shows about AGRO is displayed below.

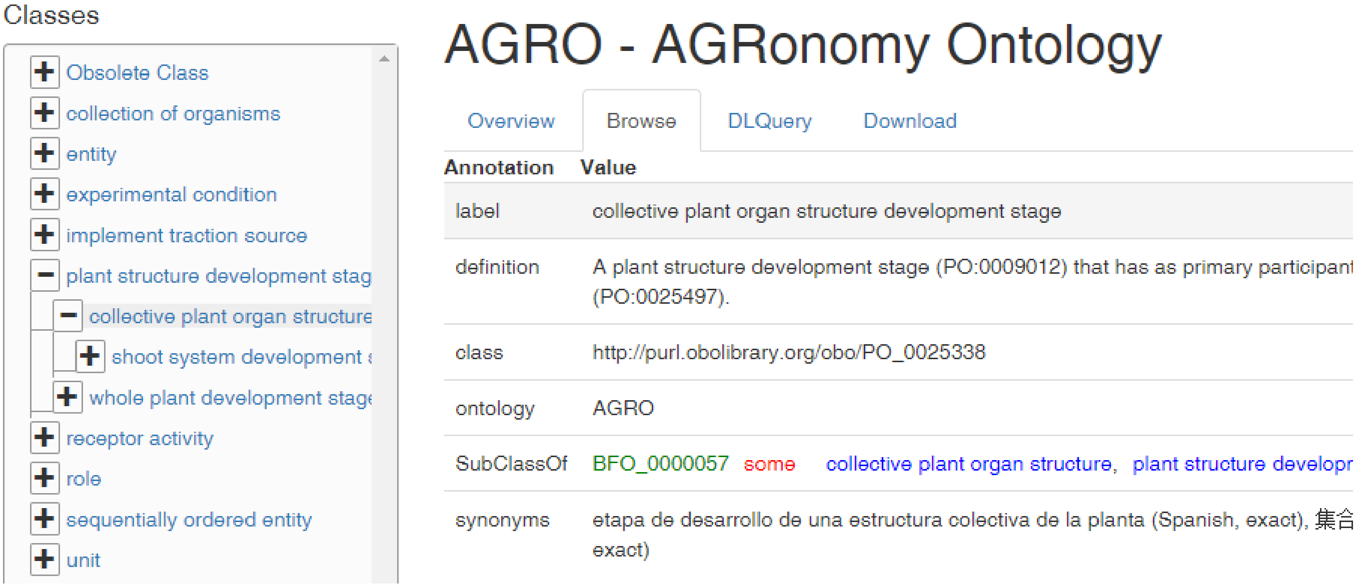

Figure 18 AberOWL Showing AGRO Classes

3 IMPROVING AGROBIO ONTOLOGIES

While researching the available AgroBio ontologies, we have observed a number of defects that could be improved to make the ontologies easier to work with, better structured, and improve their quality in general. The project consortium has contacted Bioversity International to establish a liaison for the purpose of submitting AgroBio ontology bug reports and improving the ontologies. We expect this work to continue throughout the project.

3.1 ONTOLOGY NAMING, READABILITY, PUNNING

- In many cases, the ontology namespace and ontology file location differ significantly, eg po: http://purl.obolibrary.org/obo/po# vs http://purl.obolibrary.org/obo/po.owl. This contradicts Linked Data principles, where the ontology namespace should resolve and return ontology data.
- Classes, properties and even some ontology files use numeric codes rather than English names. This makes it harder to discover and understand terms and ontologies, and therefore it is necessary to have some search/browse interface to use them effectively.
- Even rdfs:label (which is supposed to be a human-readable string) often uses unreadable abbreviations. E.g. CO_322:0001093 "EWid_M_mm" requires investigation to find out that it is related to terms
    - CO_322:0001091 "Ear width" (CO:acronym "EWid"),
    - CO_322:0001092 "EWid - Measurement", and
    - CO_322:0000206 "mm" (millimeters)
- Some local names use slash make, which makes them unsuitable for use as local names of prefixed URLs, e.g. CO_322:0000320/2 (invalid local name) is value 3="21-30% dead leaf area" of CO_322:0000320 "0-10 Senescence scale". It would be better to use another separator, e.g. CO_322:0000320_2 (valid local name)
- Using spaces in URLs results in escaping them as %20, and unsuitable for use as local names of prefixed URLs, e.g. http://www.cropontology.org/rdf/CO_356:Biotic stress
- Many entities are declared both skos: Concept, owl:NamedIndividual, owl:Class and connected by both rdfs:subClassOf and skos:broaderTransitive. E.g. for the above example CO_322:0000320/2 (value within a scale), both the value and scale are represented in this way. Although this may make for more convenient hierarchical browsing, it is not proper ontological modelling. It represents heavy punning[1] and makes OWL inference impossible.
- This also leads to redundant expression of class relations, using both owl:Restriction and a direct property e.g.

```
CO_322:0000880
    CO:variable_of      CO_322:0000132;
    rdfs:subClassOf     [ a             owl:Restriction ;
              owl:onProperty    CO:variable_of ;
              owl:someValuesFrom  CO_322:0000132 ] ;
```

- The property naming convention to start with lowercase is not always followed, e.g.

```
po:Tomato rdfs:subPropertyOf oboInOwl:SubsetProperty "Term used for tomato"
```

3.2 PREFIX PROBLEMS

- Several ontology files use empty prefixes. This is a bad practice since such a prefix cannot be distinguished from other empty prefixes

[1] https://www.w3.org/2007/OWL/wiki/Punning

- Examples of improper prefixes: rdf1: where CO: would be better (maize.owl):

```
@prefix rdf1:  <http://www.cropontology.org/rdf/>
```

- Examples of invalid prefixes (to.owl):

```
@prefix obo: <http://www.geneontology.org/formats/oboInOwl#http://purl.obolibrary.org/obo/> .
```

  - This also obscures the canonical obo: prefix, this another one needs to be used

```
@prefix obo1:  <http://purl.obolibrary.org/obo/> .
```

- Different namespaces are used for the same ontology, e.g. (to.ttl)

```
@prefix to:   <http://purl.obolibrary.org/obo/to#> .
```

  - is used only for a few meta-terms, e.g.:

```
obo:TO_0000807 oboInOwl:inSubset to:Allium_porrum
```

  - Most TO terms are defined e.g. as obo:TO_0000807, so could use:

```
@prefix TO:  <http://purl.obolibrary.org/obo/TO_> .
```

- Similarly, different prefixes are used for the same ontology (po.ttl):

```
obo:PO_0006440 # (class)
po:Angiosperm, po:derives_by_manipulation_from, po:Tomato # (properties)
```

- We define similar prefixes in upper and lower-case to account for these problems (agro-edit.ttl):

```
@prefix UO:     <http://purl.obolibrary.org/obo/UO_>.
@prefix uo:     <http://purl.obolibrary.org/obo/uo#>.
```

- CO_356 (Vitis) doesn't use any prefixes

3.3 SPECIFIC ONTOLOGY NOTES

In the rest of the section we include notes about some specific ontologies.

3.3.1 AGRO

- Source: https://github.com/AgriculturalSemantics/agro
- This is an alpha version, no official release yet
- https://github.com/AgriculturalSemantics/agro/blob/master/src/referenceMaterial/AgrO variables.xlsx could be useful for understanding the ontology. E.g. "Soil variables" has this info:

Table 3 AGRO Variables Definition excel

Variable name	SoiEle_No contact _mS/meter
Parameter	Soil electrical conductivity
Entity	Soil
Attribute	Electrical conductivity
Parameter synonyms	EC

Parameter abbreviation Other suggestion	SoiEle
Parameter description	Soil electrical conductivity is the ability of soil to conduct electrical current.
Parameter description source	http://ohioline.osu.edu/aex-fact/0565.html
Parameter class	Soil variable
Method abbreviation	No contact
Method name	No contact method
Tool / procedure	A non contact sensor works on the principle of Electromagnetic Induction (EMI). EMI does not contact the soil surface directly. The instrument is composed of a transmitter and a receiver coil usually installed at opposite ends of a non-conductive bar located at opposite ends of the instrument.
Method class: Measurement, Counting, Estimation, Computation, Observation	Measurement
Method reference	http://ohioline.osu.edu/aex-fact/0565.html
Scale abbreviation	mS/meter
Scale name	mS/meter
Scale class	Numerical

- Unfortunately, this information is neither in agro.owl nor agro-edit.owl
- Scale class: Numerical, Nominal, Ordinal, Text, Code, Time, Duration
- https://github.com/AgriculturalSemantics/agro/blob/master/src/ontology/agro.obo is empty
- https://github.com/AgriculturalSemantics/agro/blob/master/src/ontology/agro.owl has an invalid URL, as reported by Jena RIOT:

```
riot --formatted=turtle agro.owl  1>agro.ttl
10:51:21 WARN  riot            :: [line: 10060, col: 83]
 {W107} Bad URI: <http:://en.wikipedia.org/wiki/Mimicry>
 Code: 57/REQUIRED_COMPONENT_MISSING in HOST: A component that is required by the scheme is missing.
```

- This ontology defines many terms in other namespaces ("namespace hijacking"), e.g. UO, RO, etc:

```
obo:UO_0000184  a       owl:Class ;
    rdfs:label        "kilogram per meter" ;
```

3.3.2 AGRO-edit

- This is a new version in development: https://github.com/AgriculturalSemantics/agro/blob/master/src/ontology/agro-edit.owl
- It is in OWL Functional Notation, unlike agro.owl, which is RDF/XML but uses wrong file extension
- We tried to use http://mowl-power.cs.man.ac.uk:8080/converter to convert it.
  - Failed because of missing import http://purl.obolibrary.org/obo/agro/imports/po_import.owl https://raw.githubusercontent.com/AgriculturalSemantics/agro/master/imports/po_import.owl
  - This alternative works ok: https://github.com/AgriculturalSemantics/agro/raw/master/src/ontology/imports/po_import.owl
  - This "parallel" import works ok: https://raw.githubusercontent.com/AgriculturalSemantics/agro/master/imports/chebi_import.owl
  - An import of the same name (but different content) appears as:
    - https://github.com/FoodOntology/foodon/raw/master/imports/po_import.owl

  - http://www.geneontology.org/ontology/imports/po_import.owl,
  - http://purl.obolibrary.org/obo/envo/imports/po_import.owl,
  - http://purl.obolibrary.org/obo/go/extensions/po_import.owl
    http://snapshot.geneontology.org/ontology/extensions/po_import.owl (temporary failure "not found")
- Opening agro-edit.owl with Protege gives this error in OWLFunctionalSyntaxOWLParser:

```
Encountered " <ERROR> "< "" at line 7, column 1.
Was expecting:
   "Ontology" ...
   (Line 0)
```

- Opening http://purl.obolibrary.org/obo/agro-edit.owl gives no errors, but loads no ontology either
- Opening the local file AGRO-edit.owl was successful
- The project uses simple code generation with Python (called "quality patterns", because they guarantee a number of terms are generated consistently). E.g. qualityHier_2Epattern.txt has rows like this (but this particular term is not emitted in AGRO-edit.owl):

iri	iri label	entity1	entity1 label	entity2	entity2 label	attribute	attribute label	synonym	definition
AGRO_ 2000001	soil water content	ENVO_ 00001998	soil	CHEBI_ 46629	water	PATO_ 0000025	content	Above ground residue moisture	Moisture concentration of the above ground residue

3.3.3 AT

- RDF at http://data.ifpri.org/lod/at.owl, documentation at http://data.ifpri.org/lod/at/resource/. Would be better to serve both from the same URL using content negotiation
- wrong URL (extraneous #): http://data.ifpri.org/lod/at/resource/#Hybrid_maize_variety_7
- some bad namespaces, e.g.

```
@prefix j.0:   <http://purl.org/dc/terms/> . # should be dct:
```

  - Some u   nfinished individuals, e.g. (name_ what?)

```
AT:name_  a          AT:Hybrid_guinea-type_sorghum_variety ;
    AT:hasTargetCrop  crop:Sorghum .
```

  - Some non-conformance to naming conventions, e.g.

```
AT:organization       a  owl:Class . # should be capitalized
AT:rhizobial_inoculant a  owl:Class . # should be capitalized
```

- Uses a few terms from the following namespace that doesn't resolve: http://data.ifpri.org/lod/crop/
- Improperly formatted timestamp:

```
dc:date          "Jul 28, 2013 6:56:15 AM"^^xsd:dateTime ;
```

3.3.4 OEPO

- doesn't define and use these prefixes:

```
@prefix oepo:   <http://www.phenome-fppn.fr/vocabulary/2018/oepo#>.
@prefix foaf:   <http://xmlns.com/foaf/0.1/>.
```

- http://www.phenome-fppn.fr/vocabulary/2018/oepo is missing a owl:Ontology. Instead, this type (and extra metadata) is attached to a blank node
- These two nodes are disconnected, i.e. not connected to the ontology itself. Also, using owl:versionInfo for the first one is strange, it should be a version number instead of a URL:

```
<http://bioportal.bioontology.org/ontologies/URI>
   owl:versionInfo  "http://www.phenome-fppn.fr/vocabulary/2018/oepo" .
<http://bioportal.bioontology.org/ontologies/versionSubject>
   owl:versionInfo  "releases/2017-12-12" .
```

- The correct way to do this is as follows:

```
<http://www.phenome-fppn.fr/vocabulary/2018/oepo> a owl:Ontology;
 vann:preferredNamespacePrefix "oepo";
 vann:preferredNamespaceUri "http://www.phenome-fppn.fr/vocabulary/2018/oepo#";
 owl:versionInfo  "releases/2017-12-12".
```

- The ontology carries its own owl:versionInfo, which should be broken up as follows:

```
owl:versionInfo "Version 3.1";
dct:modified "2018-06-06"^^xsd:date;
dct:creator "INRA - MISTEA - LEPSE".
```

- Hijacking (redefinition) of foaf:Agent and a bunch of skos: properties
- Links are emitted as a strange mix-up of properties and URLs into a string:

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

- This should be rendered as follows (skos:exactMatch is usually used for concepts):

```
oepo:WindSensor
   owl:equivalentClass <http://purl.oclc.org/NET/ssnx/meteo/aws#WindSensor>
```

- This is even stranger because it doesn't use the semantic URL:

```
oepo:Silk rdfs:isDefinedBy "skos:exactMatch
http://www.ontobee.org/ontology/rdf/PO?iri=http://purl.obolibrary.org/obo/PO_0006488" ;
```

- Should be

```
oepo:Silk
  owl:equivalentClass <http://purl.obolibrary.org/obo/PO_0006488>
```

- This also doesn't use the semantic URL:

```
oepo:maxInclusive
  rdfs:isDefinedBy  "skos:exactMatch <https://www.w3.org/TR/xmlschema-2/#rf-maxInclusive>"
```

  - This is a URL inside some text; but the correct semantic URL is http://www.w3.org/2001/XMLSchema#maxInclusive
- Defines terms that already exist in other ontologies (namespace hijacking), e.g.:
  - oepo:Unit falls within the domain of UO, so doesn't need to be defined here
  - oepo:sfContains is copied from the GeoSPARQL ontology. Instead, the GeoSPARQL property should be used directly

- This transitive declaration makes no sense since the domain and range are disjoint. A transitive property must have the same or at least compatible domain and range. As declared, there can be no path of 2 consecutive oepo:participatesIn, so the transitive declaration is pointless

```
oepo:participatesIn a owl:TransitiveProperty;
  rdfs:domain [ a owl:Class ; owl:unionOf  ( oepo:Device oepo:ScientificObject )] ;
  rdfs:range  oepo:Experiment .
```

- Many domains and ranges are not specified, which leaves some questions, e.g.
  - What are the expected values of oepo:hasValue?
  - What is the domain of oepo:usesVector, and what vectors have to do with oepo:Device | oepo:ScientificObject?

3.3.5 CO_320 Rice

This (and other CO_* ontologies) may not be needed by the project, but we can use them as examples how to extend Vitis, and maybe we can reuse some concepts. So, we researched these ontologies and found some problems. Rice_ROOT.ttl is a tiny file that defines CO 320:ROOT as a class and concept.

- Doesn't define ontology metadata (just a blank node [a owl:Ontology])
- Redefines a number of terms from other ontologies (e.g. crop:Computation), which constitutes namespace hijacking.
- Uses both rdfs:subClassOf (which is for classes) and skos:broaderTransitive (which is for concepts, i.e. individuals). Furthermore, skos:broaderTransitive should be left to inferencing and instead skos:broader should be used in axioms, else one cannot easily find the immediate parent of a concept.

```
crop:Computation  a          owl:Class ;
      rdfs:subClassOf        crop:Method ;
      skos:broaderTransitive  crop:Method .
```

- Uses empty prefix ":" for http://www.w3.org/2002/07/owl#, which is a bad practice.
- Uses prefix "rdf1:" for http://www.cropontology.org/rdf/, which should be renamed to something more descriptive e.g. "crop:" or "co:"
- Should define and use prefix "rice:" http://www.cropontology.org/rdf/CO_320:
- Defines concepts with labels that are not comprehensible (e.g. "PanLng_MatAv_UPOV1to3"). It takes some investigation to find http://test.planteome.org/amigo/term/CO_320:0000824, where on the Graph or Tree View we can see this is a particular "rice panicle length".
- Uses some URLs with space in them e.g. rice:Biotic%20stress: bad practice, because a local name cannot include such space. Using an underscore (rice:Biotic_stress) is better
- Uses some value URLs with slash, for which the rice: prefix cannot be used, e.g. rice:00000321/1 is value "1= Strong no bending" of variable rice:00000321 "Culm strength scale SES".

3.3.6 CO_356 Vitis

Grape Ontology including OIV and bioversity descriptors. Notes:

- Created by INRA July 2017
- Homepage (curation tool) http://www.cropontology.org/ontology/CO_356/Vitis.
- The OBO file is quite shorter and easier to read http://www.cropontology.org/obo/CO_356
- Often cites this reference: Liste_des_descripteus_OIV_pour_les_varietes_et_especes_de_vitis__2e_edition_5langues_04_2008.pdf

- Search (EBI) https://www.ebi.ac.uk/ols/search?ontology=co_356
- The structure is quite simple. It defines traits, methods, scales.

Problems:

- Download as Trait Dictionary returns Server Error
- A number of incomplete/undefined terms "name: No method name found"
- E.g. CO_356:0000309
- E.g. CO_356:0000379 "No method name found"
- Is CO_356:1000215 measured in **grams** (as suggested by its name "SBER_W_g") or **milligrams** (as suggested by its relation to CO_356:4000018 "mg")?
- Uses invalid property rdfs:subProperty (it's rdfs:subProperty**Of**)
- Invalidly declares several properties (CO:method_of, CO:scale_of and CO:variable_of) as **rdfs:subPropertyOf** owl:ObjectProperty: should be **rdf:type**. Also, this constitutes namespace hijacking because the properties are defined in CO.
- CO_356:4000028 "S1_5_by2" is CO:scale_of a number of traits. It's invalidly declared a restriction owl:onProperty CO:scale_of with owl:someValuesFrom each of these traits. This means that every instance of the scale "S1_5_by2" must have links CO:scale_of to each of these traits, or else it cannot be classified with the given class. This contradicts the open world assumption, since we may have no data about some of them.
- Some terms from the OBO format are missing in the NTriples format, e.g. scale values:

```
[Term]
id: CO_356:4000033/1
name: undefined
namespace: VitisScale
synonym: "3-5-7" EXACT []
is_a: CO_356:4000033
```

- Similarly, there is extra info in the Vitis browser that is not represented in NTriples:

```
Lower limit 3.0
Upper limit 7.0
```

Finally, many terms required for AUA table grapes data are missing, e.g. "Vegetation" or "NDVI" finds nothing. NDVI is defined in specific sub-ontologies of the Crop Ontology (e.g. CO_322 Maize) but not in Crop Ontology itself.

3.4 SEARCHING FOR NDVI

We take one of the terms found in AUA table grapes data (section 4.2.2) as an example: Normalized Difference Vegetation Index (**NDVI**). We search for this term in various ontologies and examine its structure.

3.4.1 NDVI in Planteome Browser and LOV

"Vegetation index" auto-completes to 1 general and 3 specific terms:

- leaf area index (TO:0012001)
- maize normalized difference vegetation index trait (CO_322:0000132). We examine this one below
- wheat canopy normalized difference vegetation index trait (CO_321:0000301)
- wheat canopy simple ratio trait (CO_321:0000206)
- Searching in Linked Open Vocabularies (LOV), we don't find anything relevant for Vegetation:

- edac:Vegetation from ELSEweb is just a class (subclass of edac:EcologicalCommunity).
- ONTO's Proton ontology has a few types of vegetation areas, eg pext:Grassland

LOV is a widely-used ontology index and the AgroBi community should advertise its ontologies in LOV.

3.4.2 NDVI in CO_322 Maize

CO_322 Maize owl includes some terms for NDVI (see CO_322:0000880 browse neighborhood).
NDVI is represented as follows:

```
CO_322:0000132
    a                 skos:Concept , owl:NamedIndividual , owl:Class ;
    rdfs:label          "Normalized difference vegetation index"@en ;
    rdfs:subClassOf       CO_322:Physiological%20traits ;
    CO:acronym           "NDVI"@en ;
    skos:broaderTransitive  CO_322:Physiological%20traits ;
    skos:definition       "Canopy normalized difference vegetation index."@en ;
    skos:prefLabel        "Normalized difference vegetation index"@en .
CO_322:0000361
    a                 owl:NamedIndividual , owl:Class , skos:Concept ;
    rdfs:label          "NDVI - Measurement"@en ;
    rdfs:subClassOf       CO:Measurement ;
    rdfs:subClassOf       [ a              owl:Restriction ;
                      owl:onProperty     CO:method_of ;
                      owl:someValuesFrom  CO_322:0000132 ] ;
    CO:method_of         CO_322:0000132 ; # NDVI trait
    skos:broaderTransitive  CO:Measurement ;
    skos:prefLabel        "NDVI - Measurement"@en .
CO_322:0000372
    a                 skos:Concept , owl:NamedIndividual , owl:Class ;
    rdfs:label          "index"@en ;
    rdfs:subClassOf       CO:Numerical ;
    rdfs:subClassOf       [ a              owl:Restriction ;
                      owl:onProperty     CO:scale_of ;
                      owl:someValuesFrom  CO_322:0000361 ] ;
    CO:scale_of          CO_322:0000361 ;
    skos:broaderTransitive  CO:Numerical ;
    skos:prefLabel        "index"@en .
CO_322:0000880
    a                 owl:Class , skos:Concept , owl:NamedIndividual ;
    rdfs:label          "NDVI_M_idx"@en ;
    rdfs:subClassOf       CO:Variable ;
    rdfs:subClassOf       [ a              owl:Restriction ;
                      owl:onProperty     CO:variable_of ;
                      owl:someValuesFrom  CO_322:0000132 ]; # maize NDVI trait
    rdfs:subClassOf       [ a              owl:Restriction ;
                      owl:onProperty     CO:variable_of ;
                      owl:someValuesFrom  CO_322:0000372 ]; # maize index scale
    rdfs:subClassOf       [ a              owl:Restriction ;
                      owl:onProperty     CO:variable_of ;
                      owl:someValuesFrom  CO_322:0000361 ]; # NDVI meas. method
    CO:variable_of        CO_322:0000372 , CO_322:0000132 , CO_322:0000361 ;
    skos:broaderTransitive  CO:Variable ;
    skos:prefLabel        "NDVI_M_idx"@en .
```

CO:Variable ties up a specific trait (NDVI), scientific method (Measurement) and scale/unit of measure (Index).

- The representation using both owl:Restriction and direct property (CO:variable_of) is redundant.

- The terms are declared both classes (owl:Class) and individuals (skos:Concept , owl:NamedIndividual), which represents undesirable punning.
- skos:broaderTransitive is used instead of skos:broader
- The representation using direct properties can be illustrated as follows:

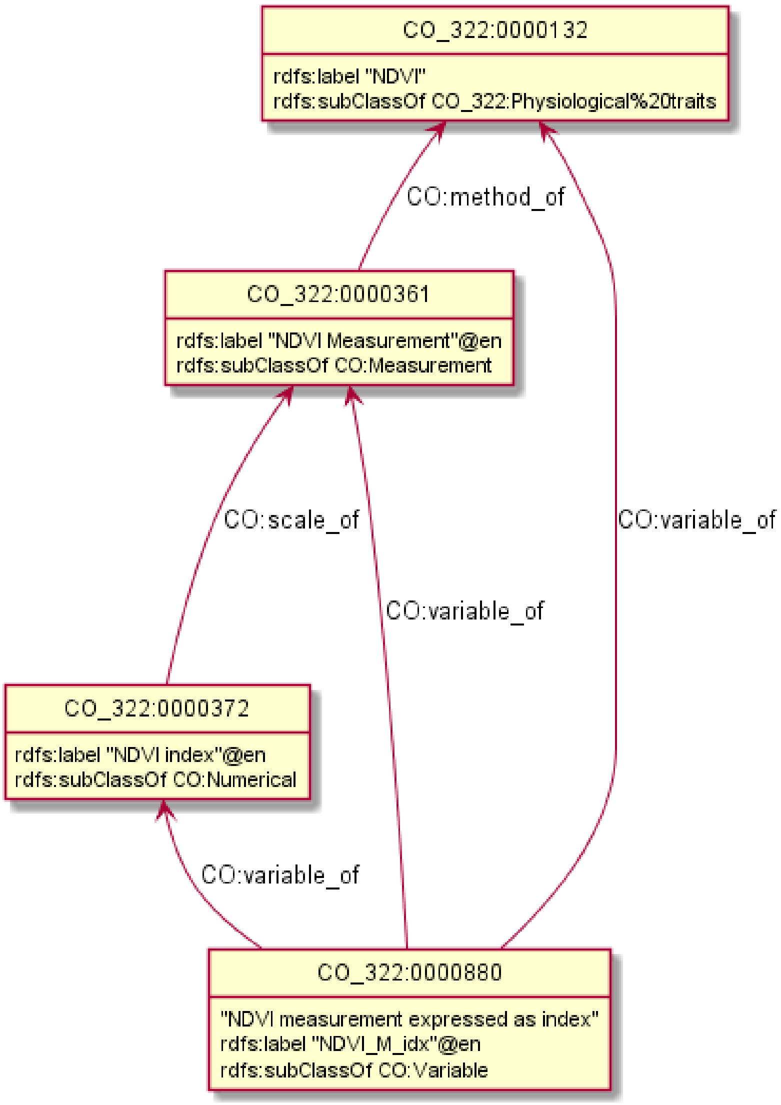

Figure 19 NDVI Representation in Maize Ontology

Problems:

- NDVI applies to all kinds of crops, so it should be in the Crop Ontology not in the Maize ontology. We believe that NDVI is defined the same for Maize and other crops.
- There is no relation to a more generic trait in CO that would apply to all crops.
- It is unclear whether there is a specific "NDVI Index" scale, and what are its values
- Using the same property CO:variable_of to connect a variable to its constituents (scale, method and trait) makes it harder to access these 3 constituents separately.

4 SPECIFIC PROJECT DATA

This section introduces data, processing requirements and data access requirements that are specific to the project.

4.1 COMPETENCE QUESTIONS

Developing semantic models or ontologies of some domain hinges on several aspects:

- 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 over-abundance of AgroBio ontologies, the latter aspect is crucial in order to keep the modelling effort focused. It should drive the following 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 modelling)
- Defining data tasks: conversion, clean-up/filtering, discretization…
- Creating sample queries to help data consumers

4.1.1 Data Domains

Data Domains defines the sort of data that we need to represent.

- Observations: when (timestamp), where (geo-reference), what (measure, dimension, attribute, and observation)
- Estates and plots, including geospatial data
- Measurement equipment
- Experiments
- Static nomenclature data, e.g.: varieties, types of measurement, etc
- Photos and other images

4.1.2 Data Questions

We have defined some draft competence questions that still need to be elaborated and validated by the partners and uses cases, to ensure they indeed are valuable research questions. The current set of questions (and elaborations for some of them) are:

- Can I retrieve the sub-plots for a given plot?
  - What's the hierarchy? Estate>Plot>Subplot?
  - Do we need/have GeoSPARQL regions for these plots? At what level?
- Which varieties are cultivated in a given plot?
- Can I retrieve weather data for a given plot?
- Which varieties are cultivated in a soil with certain characteristics?
  - How many characteristics are relevant? 10, 100, 500?
  - How are these characteristics grouped?
  - Is it meaningful to know just a few of them, or do you need to know all of them?
  - To select the optimal variety, we guess that not only the soil, but also the weather, precipitation patterns and elevation are important?

  - Will the answer be a sort of decision tree?
- Can I retrieve the origin locale for a given test sample?
  - Most probably, if we can't localize a sample, it is useless. Clarifications:
  - Does sample mean observation, or actual specimen/soil sample?
  - Does locale mean latitude/longitude/elevation? Or can it also mean specialized context, e.g. depth of a soil measurement?
  - Is localization qualifier data important (e.g. satellite number, quality of reception)?
- Can I retrieve images of a plot from which a sample was taken, at the time of collection?
  - Do we need photos of the crop at the actual time of sample taking, or only of the plot?
- Can I retrieve historical yield results for a plot (providing a timestamp)?
- Can I retrieve historical weather data for a plot (providing a timestamp)?
- Find under-performing land plots
- Is there correlation between soil conductivity and vegetation?

4.2 PARTNER AGROBIO DATA

So far, we have collected the following kinds of data from consortium partners.

4.2.1 INRA Semantic Data

INRA has submitted some sample semantic data in Github folders data/INRA/data[345].
data3 and data4 are illustrated as follows. INRA data is the top 4 nodes, and the bottom 4 nodes are from the Vitis ontology.

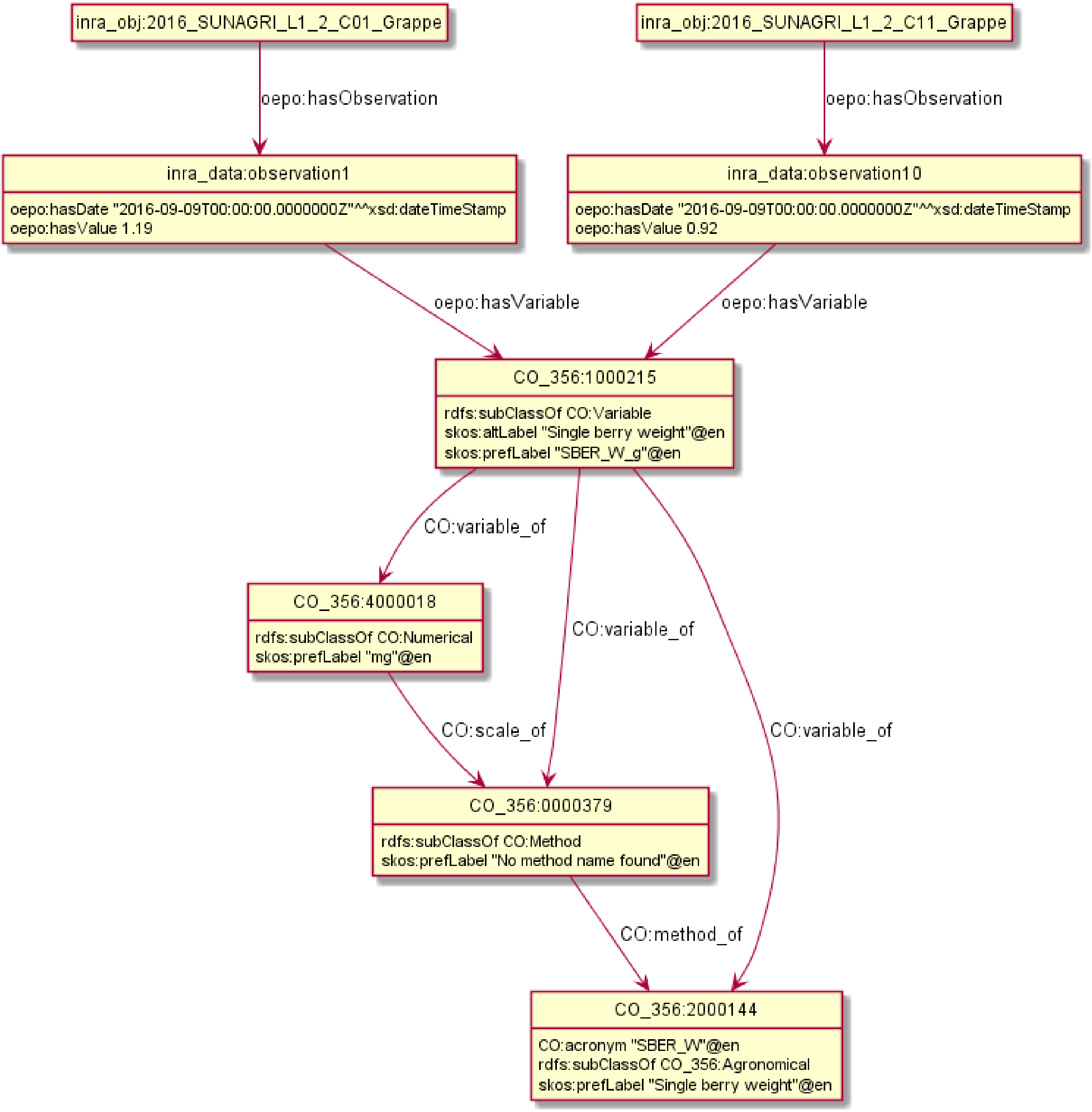

Figure 20 INRA Semantic Data

As we can see, the data consists of observations, in this illustration "single berry weight". Datasets data3 and data4 have the following problems that will be fixed by the consortium.

- Should define and use prefixes
- This is invalid datatype, should be xsd:dateTimeStamp. Alternatively, don't pad with a fake time of "0"

```
"2016-09-09T00:00:00.0000000Z"^^xsd:date
```

- http://www.cropontology.org/ontology/CO_356/Vitis#1000215 uses wrong URL, should be http://www.cropontology.org/rdf/CO_356:1000215

- http://vinnotec.supagro.inra.fr/public/Pr/data/observation1 etc are missing rdf:type
- The observed entities, e.g. http://vinnotec.supagro.inra.fr/public/Pr/2016_SUNAGRI_L1_2_C01_Grappe, are not defined in these files
- data5 includes a number of observation files, as follows:

- 2016vendanges_transf_parsed.ttl: Harvest observations: inra_onto:Poidsvendangegpesee (grams harvested).
- ComposantesGrappe_transf_parsed.ttl: Observations: inra_onto:Nbbaiescomptage number of counting bays?
- ComposantesVendanges_transf_parsed.ttl: Observations: inra_onto:Nbgrappescomptage number of counted clusters?
- fieldsLocalisationPR_parsed.ttl: plot geo-references (polygons), uses the GeoSPARQL ontology.
- FinFermentationsAlcoolique_transf_parsed.ttl
- INRA_variables.ttl: Variable definitions
- Maturite_transf_parsed.ttl
- MaturiteAnthocyanes_transf_parsed.ttl
- MaturiteJus_transf_parsed.ttl
- MaturiteSunAgri2B_transf_parsed.ttl
- must_transf_parsed.ttl: Observations: inra_onto:Sucrestotaux.brixrefractometrie Total sugars (BRIX refractometry)
- Suivifermentations_transf_parsed.ttl: Follow-up fermentations of ofpe:IntermediateProduct: observations of "Glucose/fructose g/l sequential enzymatic".

We have examined these files and made a number of recommendations, see google document README (or README.html). Often the same error applies to several terms in the same file, or to several files. E.g. the inapplicability of dct:created to time:Instant is reported for the first observation file 2016vendanges_transf_parsed.ttl but applies to all observation files.

- **Turtle prefix format.** The files use the SPARQL syntax for prefixes

```
PREFIX inra_obj: <http://vinnotec.supagro.inra.fr/public/Pr/>
```

  - While this is not an error (Turtle 1.1 supports this syntax), the older syntax supports wider interoperability:

```
@prefix inra_obj: <http://vinnotec.supagro.inra.fr/public/Pr/> .
```

- **Check against prefixes.ttl.** Use exactly the same prefixes as defined in prefixes.ttl. Consult http://prefix.cc for the most popular prefixes to use, and add to prefixes.ttl as needed.
  - Use dct: not dcterms: for DC Terms: both are valid, but the former is more popular
  - Use geo: not gsp: for GeoSPARQL, the former is a lot more popular
- **Namespaces are not suggestive.** These namespaces do not suggest they hold time and observations respectively:

```
PREFIX context: <http://www.phenome-fppn.fr/m3p/eventInsertion_ARCH2017-03-30>
PREFIX inra_data: <http://vinnotec.supagro.inra.fr/public/Pr/data/>
```

- **URLs should be resolvable.** These files use the following INRA ontologies/resources. The URLs don't resolve, and return error "Veuillez vous connecter pour avoir accès à cette page". The project should publish the data in proper semantic format, and the URLs should become resolvable.

```
inra_obj: <http://vinnotec.supagro.inra.fr/public/Pr/>
inra_data: <http://vinnotec.supagro.inra.fr/public/Pr/data/>
inra_agent: <http://vinnotec.supagro.inra.fr/public/Pr/agent/>
```

```
inra_code:  <http://vinnotec.supagro.inra.fr/public/Pr/code/>
inra_onto:  <http://vinnotec.supagro.inra.fr/public/Pr/onto/>
```

- **syntax error (unquoted string)**

```
[line: 183, col: 24] Unrecognized: divers
inra_obj:JARDIN-AMPELO divers rouge rdf:type aeo:Plot ;
```

- **dct:created is inappropriate**: one can't "create" a time instant (it just exists), so dcterms:created is inappropraite. To express when an event was converted (vs occurred), we could use the PROV ontology.

```
context:instant_e1ba2667-2a37-4a42-b157-7ac07bfc458e rdf:type time:Instant ;
  time:inXSDDateTimeStamp "2016-08-24T12:00:00+01:00"^^xsd:dateTimeStamp ;
  dcterms:created "2018-07-12T18:52:00.012981"^^xsd:dateTime .
```

- **aeo:involvedIn is inappropriate**. Plots are **part of** Lots, they are not **involved in** lots. aeo:involvedIn is defined as "AgriExperiment involves different instances of AgriActivity and AgriEntity")

```
inra_obj:81-CHARDONNAY rdf:type aeo:Plot ;
 aeo:involvedIn inra_code:Lot_FV-2016-002 ;
```

- **Class vs Property**. This is a class not a property, so it can't be used like this. (In general, I notice that all AgroBio ontologies have lots of classes but few properties).

```
ofpe:Operator inra_agent:fabien.robert ;
```

- **rdf:value?** I can't verify whether oepo:Observation can take rdf:value because OEPO doesn't define this. Using rdf:value this way could be ok, but we should specify it with an RDF Shape.
- **invalid DateTimeStamp**, as reported by Jena RIOT.

```
[line: 16, col: 28] Lexical form '09/09/16' not valid for datatype xsd:DateTimeStamp
```

- **missing rdf:value**. Jena RIOT reports an error, which is caused by a missing rdf:value in the observation.

```
[line: 491, col: 47] Triples not terminated by DOT
inra_data:4e1956e2-eceb-477f-97a4-d22a919970b1 rdf:type oepo:Observation ;
 time:hasTime context:instant_39dec42b-9d84-4269-96f6-289d0d0ee782 ;
 oepo:hasVariable inra_onto:Nbbaiescomptage ;
```

- **Indicate grape variety**. Plots don't seem to indicate the grape variety, except in the URL, but a URL should be interpreted as opaque and not information-bearing.

```
inra_obj:22-SYRAH rdf:type aeo:Plot .
inra_obj:68-COLLECTION-BLANCS rdf:type aeo:Plot .
```

- **Use QUDT**. Plot areas are described using DBpedia and the Telegraphis Quantity ontology (which returns 404 Not Found). However, we better use the QUDT ontology that is more popular and has a full complement of SI and other kinds of units, including expression of units in terms of fundamental quantities (time, mass, length, etc) and conversion factors between units.

```
PREFIX dbo:      <http://dbpedia.org/ontology/>
PREFIX quty:     <http://www.telegraphis.net/ontology/measurement/quantity#>

inra_obj:81-CHARDONNAY rdf:type aeo:Plot ;
  oepo:hasObservation inra_data:6870097e-13b9-4179-83c3-78450c0bb8ce .
inra_obj:81-CHARDONNAY rdf:type aeo:Plot ;
  quty:area "1.20600"^^dbo:hectare .
```

- **Fix polygon geometry**. Plot polygons as defined include just 4 coordinates. Even for a simple box you need 4 corners, i.e. 8 coordinates. Coordinates should be +-180 degrees longitude and +-90 degrees latitude, but these are very big numbers. There are two pairs of the same number, but these should be "lat lon" pairs.

```
inra_obj:81-CHARDONNAY gsp:hasGeometry inra_gis:polygon_81-CHARDONNAY .
```

```
inra_obj:81-CHARDONNAY rdf:type aeo:Plot ;
gsp:asWKT "POLYGON ((710743.61182814 710743.61182814, 6226766.01933858 6226766.01933858 ))"^^gsp:wktLiteral .
```

- After coordinates are fixed, we need to check them for validity:
  - Order of latitude/longitude
  - That it indicates a place in France
  - That the given area in hectares corresponds to the polygon's area
- **gsp:Polygon vs gsp:Geometry**. There's no class gsp:Polygon. Use gsp:Geometry instead
- **Declare geo:Feature**. geo:hasGeometry has domain geo:Feature, so it should be declared, e.g. as

```
inra_obj:22-SYRAH rdf:type aeo:Plot, geo:Feature.
```

- **Namespace hijacking**. Don't define terms of other ontologies:

```
CO:variable_of  rdfs:subProperty_of  skos:related ;
     rdf:type           owl:ObjectProperty .
```

- **Use English class names**. To make ontologies that are more easily understood and reusable, we should use English

```
inra_onto:Poidsvendangegpesee # weight as measured at vine picking
```

- **Define labels**. E.g. inra_onto:Glucose.fructoseg.lsequentiel.enzymatique needs a label such as "Glucose/fructose g/l sequential enzymatic"
- **Can't use CO_UO "gram"**. Checking whether inra_onto:Poidsvendangegpesee defines everything required to interpret the number, we find the following data.

```
inra_onto:Poidsvendangegpesee CO:variable_of CO_356:2000168 , CO_UO:0000021 , MMO:0000157 .
CO_356:2000168 rdfs:label "Yield"@en .
CO_UO:0000021 rdfs:label "g"@en; CO:scale_of CO_357:2000105 .
CO_357:2000105 rdfs:label "Ratio shoot root protocol"@en .
MMO:0000157 rdfs:label "digital scale post excision weight measurement" .
```

  - CO_UO:0000021 "gram" is defined as a scale of "ratio shoot root" (some Woody Plant feature), so it cannot be used for grapes. This is yet another example of over-specialization (improper lack of abstraction) in AgroBio ontologies.
  - Note: one can get the whole CO_UO from neither http://www.cropontology.org/rdf/UO: nor http://www.cropontology.org/rdf/UO. But individual terms are returned, e.g. http://www.cropontology.org/rdf/UO:0000021 returns Turtle.
- **Missing CO_UO Term**. http://www.cropontology.org/rdf/UO:0000175 is missing: unlike the above UO:0000021, this one returns nothing.

```
inra_onto:Glucose.fructoseg.lsequentiel.enzymatique CO:variable_of
  CO_356:2000057, CO_UO:0000175, MMO:0000388 .
```

- **Reflexive subclass**. AEO defines a reflexive subclass relation (last pair in the chain below), which is implied by RDFS and is useless

aeo:Plot < aeo:CultivatedLand < aeo:Area < **aeo:AgriEntity < aeo:AgriEntity**

- **Syntax error**. The problem is missing a prefix of the subject.

```
[line: 28, col: 1 ] Broken token (newline): VIP_Sauvignon rdf:type afeo:Must ;
```

- **Syntax error**.

```
[line: 144, col: 26] Unrecognized: sec
```

- **Class vs Property**. oepo:Observation needs some link to Agent, be that Operator or Organization. But foaf:Organization is a class not a property so it can't be used like this.

```
inra_data:32757c4a-15dd-4896-a3b9-970f33e6f756 rdf:type oepo:Observation ;
  foaf:Organization inra_code:16-1841 ;
```

- **Where are inra_codes defined?** These codes are used by the data, but are not defined anywhere.

```
inra_code:Cuve_BB1010 # FinFermentationsAlcoolique_transf_parsed
inra_code:BB1010     # Suivifermentations_transf_parsed
```

- **Organization individuals.** Organization URLs (e.g. inra_code:16-1841) use some codes. These URLs should be defined as proper individuals and may be better to use some more suggestive URLs.

4.2.2 AUA Tabular Data

AUA has submitted tabular observation data (soil, plant canopies, spectral vegetation indexes) about table grapes.

- See the data in WP8/Table Grapes Pilot- AUA/Data. See Photos for some images.
- See D8.1 Piloting Plan (specifically BigDataGrapes_Piloting Plan-AUA) for descriptions of the equipment and measured indicators
- The measurements are made with 4 kinds of equipment: EM38, RapidScan, SpectroSense, Crop Circle:
  - Measurements for Soil Electrical Conductivity are taken with an **EM38** device
  - Measurements include information from plant canopies and classic spectral vegetation index data (NDVI, NDRE etc.) with **RapidScan, SpectroSense** and **Crop Circle**
- There about 10 measurements per measurement spot
- The measurements are Geo-referenced (longitude, latitude, altitude) and timestamped
- Includes 3 estates: Fasoulis, Kontogiannis, Palivou. Each estate is subdivided into a number of plots. The plots are named after:
  - Grape varieties: mavroudi, roditis, savatiano, soultanina (Kontogiannis Estate); Merlot (Palivou Estate)
  - Nearby settlements: solomos (Kontogiannis estate)
  - Names given by the owners or relative to the location: Geotrisi, IFG, Kato (Fasoulis Estate); Alekos, dipla oinopoiio, kato, mesi, pano (Palivou Estate)
- **Boundaries** and **Elevation** files give the plot spatial coordinates, e.g.: Fasoulis_RTKGPS_Boundaries.csv, Kontogiannis_RTKGPS_Boundaries.csv, Kontogianis_RTKGPS_Elevation.csv, Palivou_RTKGPS_Boundaries(all).csv, Palivou_RTKGPS_Elevation(all).csv

For example, file "5. Fasoulis_IFG_RapidScan.xlsx" includes tabular info like this (22 columns):

PLOT	NDRE	NDVI	RE	NIR	R	LATITUDE	LONGITUDE	ELEVATION	HDOP	FIXTYPE	DATE
37	0.2252	0.7376	20.836	33.084	5.132	37.81713	22.58971	291.5	2.8	GPS	5/23/2018

TIME	N	MAXNDRE	MAXNDVI	MINNDRE	MINNDVI	STDNDRE	STDNDVI	CVNDRE	CVNDVI
10:12:50	256	0.3423	0.8872	-0.3207	-0.0788	0.0784	0.1675	0.3479	0.2271

See AUA Table Grapes Data for some notes on measurement equipment and specific measurements

- EM38 measures apparent soil electrical conductivity (ECa):
  - Longitude
  - Latitude
  - CV1m: conductivity at depth 1 meter in millisiemens per metre (mS/m)
  - CV05m: conductivity at depth 0.5 meter in millisiemens per metre (mS/m)
  - Quality, Satellite, HDOP: related to the GPS signal-explained below
  - Elevation
  - Time and Date given by the GPS

- RapidScan measures Canopy characteristics and vegetation indices:
  - RE: Red-Edge spectral region (spectrum centred around 715 nm)
  - R: Red spectral region
  - NIR: Near-infrared spectral region
  - NDRE: mean value Normalized Difference Red Edge Index, defined using NIR and RE
  - NDVI: mean value Normalized Difference Vegetation Index, defined using NIR and R: (NIR-R)/(NIR+R)
  - Latitude
  - Longitude
  - Elevation
  - HDOP, FIXTYPE: related to the GPS signal-explained below
  - Date, Time
  - MAXNDRE, MAXNDVI: maximum values for NDRE and NDVI
  - MINNDRE, MINNDVI: minimum values for NDRE and NDVI
  - STNDVI, STNDRE: standard deviation for NDRE and NDVI
  - CVNDRE, CVNDVI: coefficient of variation for NDRE and NDVI
- Both equipment record a GPS and datetime fix:
  - Longitude, Latitude, (or Northing and Easting on a UTM projection ZONE 34N) Elevation
  - Time, Date
  - HDOP: horizontal dilution of precision, a factor in determining the relative accuracy of a horizontal GPS fix
  - Quality: quality of the GPS receiver (EM38 only)
  - Sat: which satellite provided the GPS fix (EM38 only)
  - PLOT: sequential measurement number in this run (RapidScan only). Note: this is **not** a plot number
- SpectroSense measures canopy characteristics and vegetation indices:
  - Context:
    - Northing, Easting: a specific way of expressing coordinates
    - Elevation
    - Satellite
    - HDOP
    - Date and Time
    - Mod: related to the GPS signal
  - Canopy characteristics:
    - REDi: Incident radiation of the red spectrum
    - REDr: Reflected radiation of the red spectrum
    - NIRi: Incident radiation of the Near-InfraRed spectrum
    - NIRr: Reflected radiation of the Near-InfraRed spectrum
  - Then we calculate the following:
    - NIR: NIRr / NIRi
    - RED: REDr / REDi
    - NDVI: Normalized Difference Vegetation Index = (NIR - RED) / (NIR + RED)
    - LAI: Leaf Area Index = 0.0148*(EXP(6.192*NDVI))
  - Optical Measurement Bands: (SF1-SF3 User definable and SF4, SF5 calculated by the sensor)
    - SF1 - channel with 670 nm (BW ±11 nm) interference filter
    - SF2 - 730 nm (BW ±10 nm) interference filter
    - SF3 - 760 nm (LWP) interference filter
    - SF4 and SF5

To tie measurements to a specific plot, geo-coordinates need localization within the plot (GeoSPARQL **within** predicate).

4.3 DATA PROCESSING REQUIREMENTS

This section outlines some specific data processing requirements to be taken into account by WP3. We expect that more requirements will be defined as the use cases progress further.

4.3.1 Data Validation and Handling

Based on the syntactic and semantic errors observed above, we have started a guideline for data validation and handling. It covers:

Started rules on:

- How to submit files. We currently use Github, which is synchronized with google drive, but should select only one of them.
- How to use and update prefixes.ttl, a common prefixes file to be used consistently by all project partners.
- How to validate RDF file syntax using Jena **riot** (and maybe Jena **eyeball**)

This will grow to a comprehensive document on semantic data handling and validation by project partners.

Future topics include:

- Semantic conversion tools
- Semantic validation. Once we decide on patterns for representing data, we plan to implement RDF Shapes and a Validation Service to ensure the quality of data collected from the partners and converted by the project.

prefixes.ttl
The project keeps a single master prefixes file: prefixes.ttl (this is currently in ontology/notes/img/, but will definitely move to a more meaningful location).

- All partners should ensure they use the same namespaces and prefixes (e.g. dct: not dcterms: for Dublin Core Terms, and geo: not gsp: for GeoSPARQL)
- Check your prefixes against prefixes.ttl: if there's a discrepancy, discuss with Ontotext
- If you need a new prefix: consult http://prefix.cc for the most popular one, add it to prefixes.ttl and commit.

As a best practice, do not include individual prefixes in Turtle files, instead always prepend prefixes.ttl. This is especially important if you exchange a large number of small/example files.

Syntax Validation

- Use RIOT (part of Apache Jena) to validate the syntax of your files, e.g.

```
riot --validate 2016vendanges_transf_parsed.ttl
```

- If you prepend prefixes to Turtle files, use the script **riotval.pl:** it prepends prefixes, calls RIOT validation, then subtracts the number of lines in prefixes.ttl from error messages.
- If you are more adventurous, also try Jena Eyeball that performs deeper validation (e.g. that unknown class/property names are not used). However, there is no Apache release of Eyeball and the code has not been updated for Jena3.

4.3.2 Data Cleaning

Use case A. Data Anomaly Detection & Classification defines some needs for data cleaning. E.g. see this row:

- Name: Eca sensing;
- Description: Georeferenced soil electrical conductivity data;
- Operations Performed: Data filtering for outliers:
- Provenance: Proximal sensors

EM38 is affected by metal pillars (poles), so soil conductivity readings near such poles make the measurement invalid. E.g. on Fasoulis_Kato_EM38_map (metal vineyard pillars).jpg, red readings show the position of pillars, and only the green readings should be retained. Readings over the value 100 should be discarded.

Another example is: **RapidScan** needs some time to establish a GPS connection. See file 6. Fasoulis_Geotrisi_RapidScan.xlsx for some examples. The following kinds of measurement should be discarded because they don't have a valid geo-reference:

- Readings with "FIXTYPE: Fix not valid" (missing geo-coordinates)
- Readings with negative ELEVATION (invalid geo-coordinates)

4.3.3 Data Discretization

In order to correlate multiple measurements made in "essentially" the same context, we need discretization:

- Spatial: geo-coordinates need Discretization (e.g. to a grid of 2x2m), and then Averaging
- Temporal: we need rules for datetime discretization, e.g. whether it is it ok to average two measurements done within a day.

4.3.4 Data Localization

To link metrics to a specific sub-plot, we may need to localize geo-coordinates within a sub-plot. Assuming that we have the sub-plot polygons, we can use the GeoSPARQL predicate **within**. Ontotext GraphDB supports a full complement of GeoSPARQL relations, using 3 different spatial relation algebras.

4.4 DATA ACCESS REQUIREMENTS

There are some impediments to effective use of semantic technologies by AgroBio researchers that we need to address (these are in addition to semantic data integration steps/challenges as outlined in sec 1.4):

- Given the huge number of AgroBio ontologies, it is hard for researchers to find and effectively apply them.
- AgroBio researchers should not be expected (in most cases) to write SPARQL: they need a simpler way to get data out of the semantic Knowledge Graph, i.e. query writing aids and visualization mechanisms.

Regarding the first challenge, we need to deploy and/or develop simple discovery tools, both at the level of ontologies and the level of individual terms. Section 2.4 describes a number of ontology portals and search tools, and sections 2.4.3 and 2.4.10 specifically describe Annotator tools that may ease the application of ontologies to data and text.

Regarding the second challenge, there are a number of applicable approaches, e.g. see Data Visualization with GraphDB and Workbench[2] for an overview.

- Build a JDBC/ODBC interface to Ontotext GraphDB to allow execution of a SPARQL query and receiving results through these interfaces. This will be very convenient for use in visualization and analytics tools, e.g. Tableau and PowerBI.
- Develop canned (predefined) queries that answer validated Competence Questions.
  - Create convenient ways to deploy such queries as REST services. See GRLC[3] (Meroño-Peñuela and Hoekstra, 2016; Meroño-Peñuela and Hoekstra, 2017) and SPARQL2Git[4] (Meroño-Peñuela and Hoekstra, 2017a)
  - Describe and parameterize the queries, so researchers can provide query inputs.
- Investigate approaches to translate natural language to SPARQL, for example:
  - Based on translation grammars, e.g. Grammatical Framework[5] (Marginean et al. 2014)
  - Based on machine learning, e.g. seq2seq neural networks (Soru et al., 2017)
- Trial and experiment with Visual Query Builders, e.g. ViziQuer (Cerans et al., 2017; Cerans and Ovcinnikova, 2012016) or SPARKLIS (Ferré, 2015). See below for two illustrations or (Soylu, 2017) for a comprehensive review of similar systems.

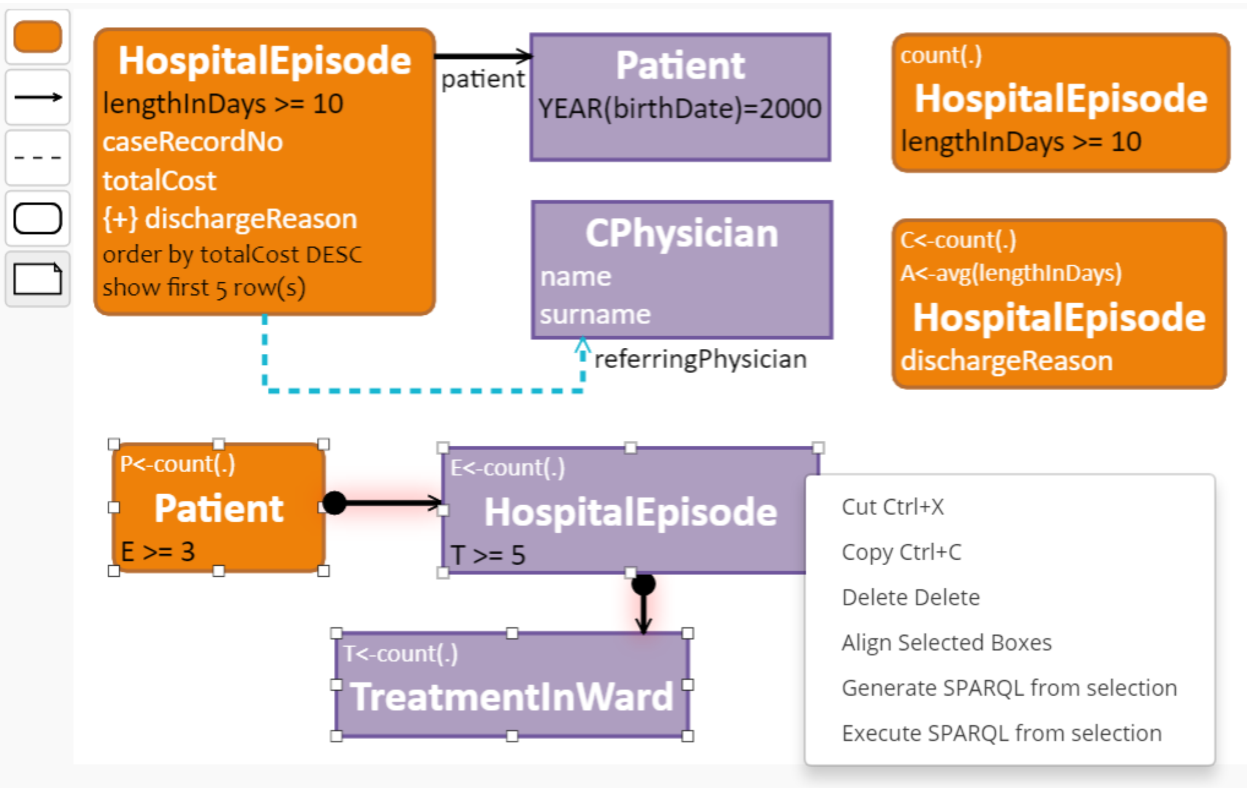

Figure 21 Sample Query in ViziQuer

[2] Data Visualization with GraphDB and Workbench: https://docs.google.com/document/d/1guwFHi9p4-ujFkrHF6dwMUZndzCmlX_gPyiBi6JlPTs/pub

[3] http://grlc.io/, source at https://github.com/CLARIAH/grlc/

[4] http://sparql2git.com, source at https://github.com/albertmeronyo/SPARQL2Git

[5] http://www.semantic-web-journal.net/content/question-answering-over-biomedical-linked-data-grammatical-framework-0

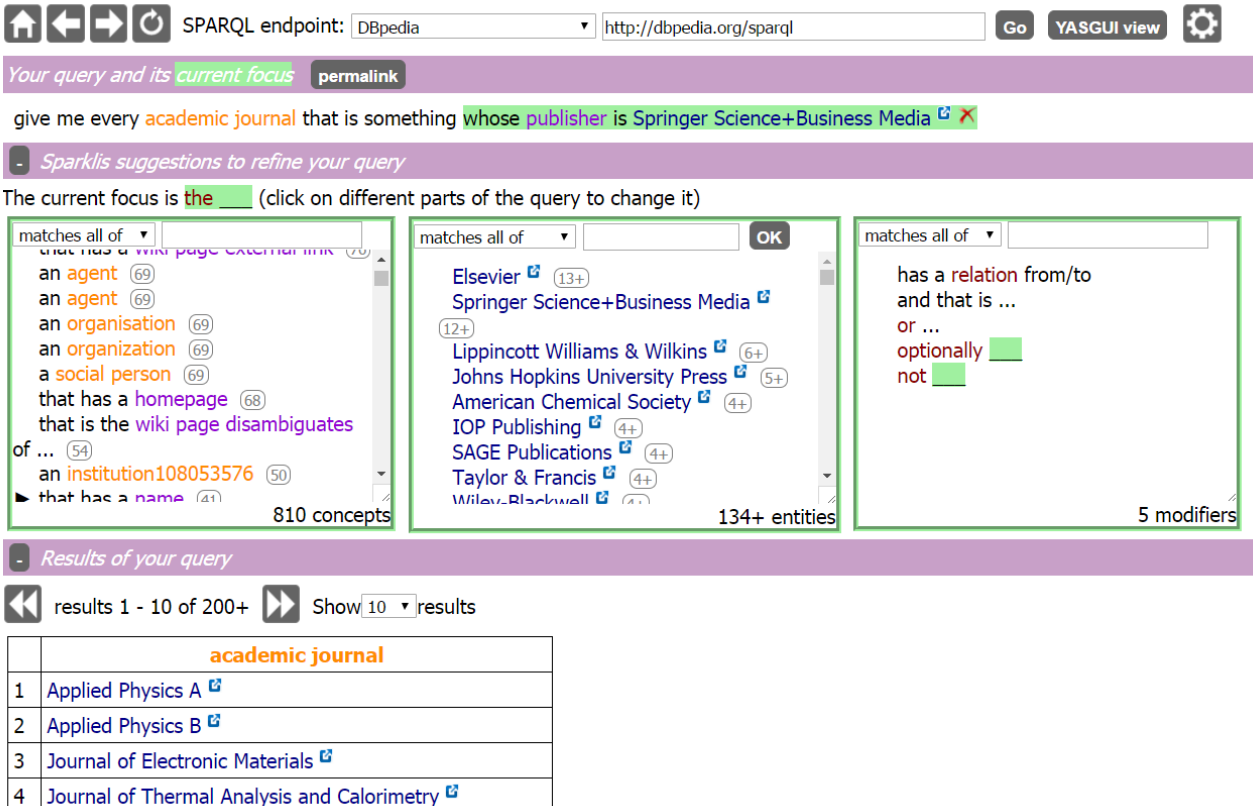

Figure 22 Sample Query in SPARKLIS

5 CONCLUSIONS

This document presented the progress on WP3 Data & Semantics Layer during the first 9 months of the BigDataGrapes project. More specifically, it has presented:

- The sort of data to represented in a semantic way
- Specific steps that we intend to follow for Semantic Data Integration
- Relevant AgroBio ontologies and problems that we have found in them
- Related work: AgroBio portals and other related tools
- Specific project data
- Specific data processing requirements
- Specific data access requirements and relevant tools and approaches

Section 1.4 outlined the steps that WP3 partners intend to follow, and the approximate point that has been reached within these steps. This section outlines the immediate next steps to be taken.

- Define data needs that will drive technical development
  - Continue the work on competence questions (section 4.1). Validate competence questions against data needs dictated by the use cases specified in task T2.1
  - Work with T2.1 to define expected data volumes and data formats. This will drive the selection of conversion tools
  - Define any data update requirements
- Define best practice data models for representing fundamental AgroBio data (see section sec 1.1 and the discussion in section 1.2)
  - Create a comprehensive example of AUA tabular data (sec 4.2.2) represented in the W3C CUBE ontology (section 1.3)
  - Discuss CUBE vs traditional AgroBio ontological representation with INRA and other project partners
  - Agree the recommended data models and document them with rdfpuml
- Continue to liaise with the AgroBio community on improving the quality of the respective ontologies (see section 3)
- Implement an example conversion of AUA data with OntoRefine[6], an ONTOTEXT tool for working with tabular data
- Elaborate requirements for semantic discovery, mapping and linking tools, using those reviewed in section 2.4 for inspiration
- Elaborate data access requirements (see section 4.4) and start work on the respective tooling

[6] http://graphdb.ontotext.com/documentation/standard/loading-data-using-ontorefine.html

REFERENCES

Cerans, K., Barzdins, J., Sostaks, A., Ovcinnikova, J., Lace, L., Grasmanis, M. and Sprogis, A. (2017). Extended UML Class Diagram Constructs for Visual SPARQL Queries in ViziQuer/web. Available online: http://ceur-ws.org/Vol-1947/paper08.pdf

Cerans, K. and Ovcinnikova, J. (2016). ViziQuer: Notation and Tool for Data Analysis SPARQL Queries. Available online: http://ceur-ws.org/Vol-1704/paper15.pdf

Ferré, S. (2015). Sparklis: An Expressive Query Builder for SPARQL Endpoints with Guidance in Natural Language. Semantic Web Journal. Available online: http://www.semantic-web-journal.net/content/sparklis-expressive-query-builder-sparql-endpoints-guidance-natural-language-0

Marginean, A., Groza, A., Slavescu, R.R., Alfred Letia., I. (2014). International Conference on Development and Application Systems, DOI: 10.1109/DAAS.2014.6842456. Available online: https://www.researchgate.net/publication/263218121_Romanian2SPARQL_A_Grammatical_Framework_approach_for_querying_Linked_Data_in_Romanian_language

Meroño-Peñuela, A., and Hoekstra, R. (2016). grlc Makes GitHub Taste Like Linked Data APIs. ESWC 2016 Satellite Events (SALAD 2016), Heraklion, Crete, Greece, May 29 – June 2, 2016, Revised Selected Papers. LNCS 9989, pp. 342-353 (2016). Available online: https://link.springer.com/chapter/10.1007/978-3-319-47602-5_48

Meroño-Peñuela, A. and Hoekstra, R. (2017). ISWC 2017, 16th International Semantic Web Conference. Lecture Notes in Computer Science, vol 10587, pp. 334-339 (2017). Available online: https://iswc2017.semanticweb.org/wp-content/uploads/papers/MainProceedings/430.pdf

Meroño-Peñuela, A. and Hoekstra, R. (2017a). SPARQL2Git: Transparent SPARQL and Linked Data API Curation via Git. Proceedings of the 14th Extended Semantic Web Conference (ESWC 2017), Poster and Demo Track. Portoroz, Slovenia, May 28th – June 1st, 2017. DOI 10.1007/978-3-319-70407-4_27. Available online: https://pdfs.semanticscholar.org/85cc/73ede853e8f7d9c1c7371c4b435a80123af3.pdf

Soru, T., Marx, E., Moussallem, D., Publio, G., Valdestilhas, A., Esteves, D. and Baron Neto., C. (2017). Posters and Demos Track of the 13th International Conference on Semantic. Amsterdam, The Netherlands, September 11-14, 2017. Available online: http://ceur-ws.org/Vol-2044/paper14/

Soylu, A., Giese, M., Jimenez-Ruiz, E., Kharlamov, E., Zheleznyakov, D. and Horrocks, I. (2017). Universal Access in the Information Society, June 2017, Volume 16, Issue 2, pp 435–467. DOI 10.1007/s10209-016-0465-0. Available online: https://link.springer.com/article/10.1007/s10209-016-0465-0