

# Semantics for Crops & Tools for Data Annotation

Elizabeth Arnaud, Bioversity International



# Data Harmonization along the data management workflow and across IT tools

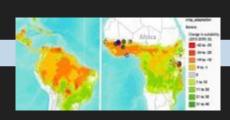
### Semantics for harmonization & interoperability



#### **Data storage**



#### **Data Analysis**



#### **Data Access**



- BMS Fieldbook
- Banana Fieldtask
- KSU Fieldbook
- AGROFIMS fieldbook
- KDSmart
- ODK surveys
- DECODING
  The Data Ecosystem

- BMS
- RTB NextGen Databases
- AGROFIMS
- Breeding4Rice
- KDExplore
- Dataverse, CKAN
- MARLO/MEL

- IBP
- HIDAP
- FarmDesign
- Crop models
- ....





# FAIR Data in Agriculture





Interoperable - The most difficult to achieve and requires ontologies/controlled vocabularies identification of key objects (Ontologies)

Use of proper concept identifiers



### What is a vocabulary?

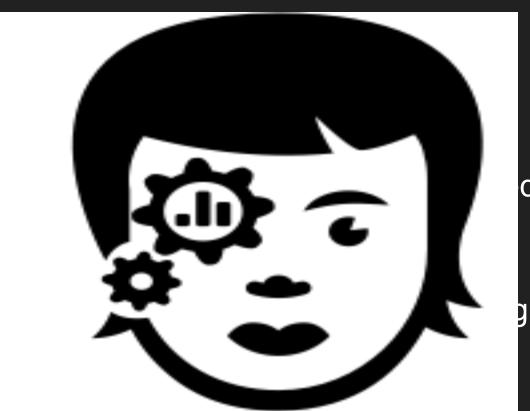
- **Controlled vocabularies:** 
  - provide a shared vocabulary for a domain (all the terms)
  - provide textual definitions that describe the intended meaning of the classes in vocabularies
  - provide standard identifiers for concepts describing a given domain
- Facilitate data publication and data access





## What is a vocabulary?

- Controlled voc
  - provide a voca
  - provide **textua**l classes in voca
  - provide standa
- Facilitate data



d meaning of the

g a given domain





## What is an ontology?

#### Ontologies:

- provide standard identifiers for classes and relations that represent the phenomena
   within a domain
- provide a vocabulary for a domain
- provide textual definitions that describe the intended meaning of the classes and relations in ontologies
- provide machine-readable axioms and definitions that enable computational access to some aspects of the meaning of classes and relations

Facilitate data integration, data access and analysis



## What is an ontology?

### Ontologies:

provide standard ide

within a domain

provide a **vocabular** 

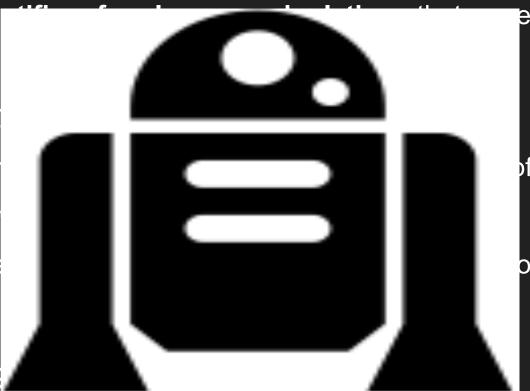
provide **textual defir** 

relations in ontologie

provide **machine-re**a

some aspects of the

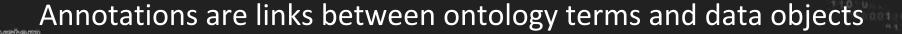
Facilitate data integ



esent the phenomena

of the classes and

omputational access to





# Harmonizing Phenotypic data





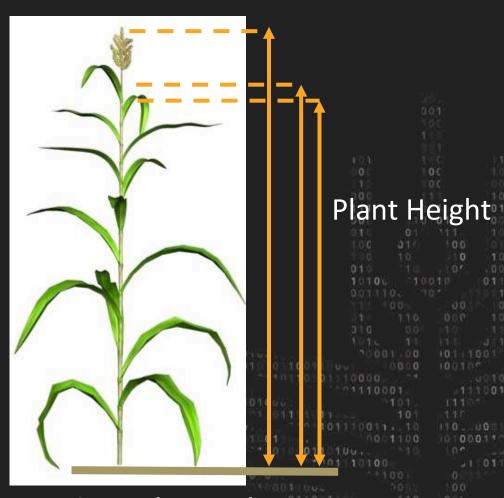


# **Crop Ontology**

# Breaking semantic barriers to data interpretation

- Breeding Management Systems
- Meta Analysis across evaluation sites
- Confusion between traits and variables
- No naming convention for variables and methods of measurement which are heterogeneous
- Trait & Variable definitions and measurements are not similar between farmers, breeders, agronomists, modelers





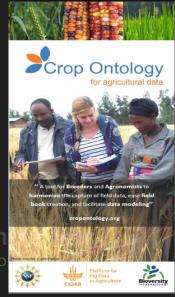
=> One trait = 4 traits...







Provides measured traits, parameters and their standard variables for the Breeding fieldbook and for data annotation in the crop databases







Integrated Breeding
Today's tools for tomorrow's cro



























### Phenotype data in breeding

Germplasm ID	PH	PH2	GCOI	ر ا	GY	
24530	80	1	2		35	
85432	120	3	4		48	
78452	95	2	4		43	
56093	100	2.	1		50	

The height of the plant

The color of the grain

Annotation must explain:

1/ What is the observation about ?

=

# TRAIT Entity + Attribute













### Phenotype data in breeding

Germplasm ID	PH	PH2	GCOL	GY
24530	80	1	2	35
85432	120	3	4	48
78452	95	2	4	43
56093	160	2	1	50

Annotation must explain:

2/ How is the trait observed?

=

**METHOD** 

With a measuring tape

Estimated visually

Calculated
Panicle weight x panicle density











### Phenotype data in breeding

Germplasm ID	PH	PH2	GCOL	GY
24530	80	1	2	35
85432	1/20	3	3	48
78452	95	2	3	43
56093	100	2	1	50

Annotation must explain:

3/ How is the trait observation expressed?

cm

1 = short (<90cm)

2 = medium (>90cm and

<110cm)

3 = tall (>110cm)

1 = white

2 = cream

3 = yellow

**SCALE** 

g/m<sup>2</sup>



# Standard Variable in Crop Ontology

Method and scales are important to capture for supporting data comparison
 & interpretation

A **Variable** is combination of

**Property (Trait) + Method + Scales/units** 

- > It annotates the actual value of the measurement
- > Has a unique name -> Proposed naming convention: P\_M\_S



Trait = Entity+ Quality
(Root) (Length)



Average root length expressed in cm

Average root length scored on a 3 level scale :

1=short (<50cm)

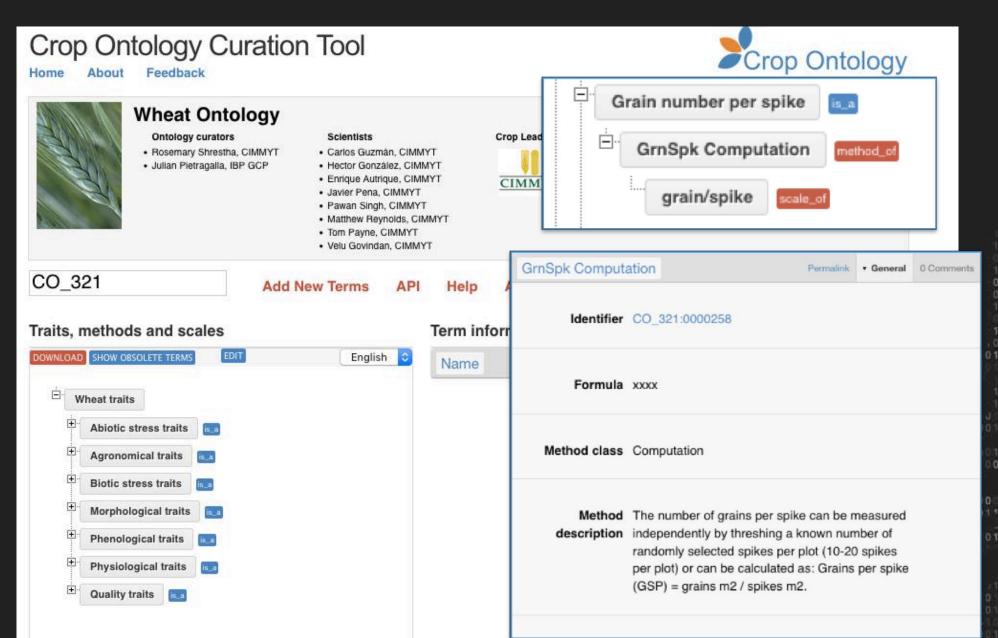
2= medium (50-100cm)

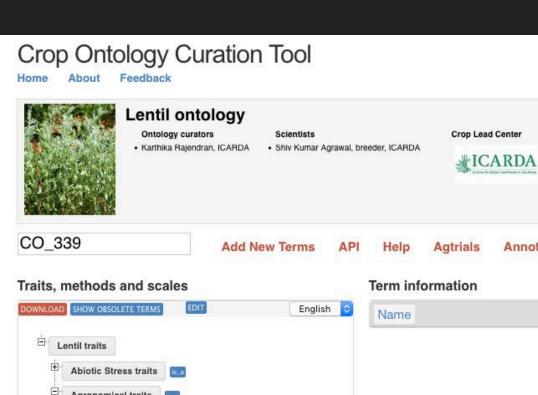
3=long (>100cm)

# **Crop Ontology Online**

- 27 species
- Download in csv, obo,
   RDF
- API
- Helpdesk
- Adding farmers preferences, sensory traits





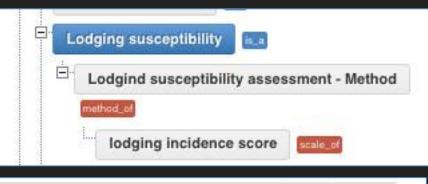


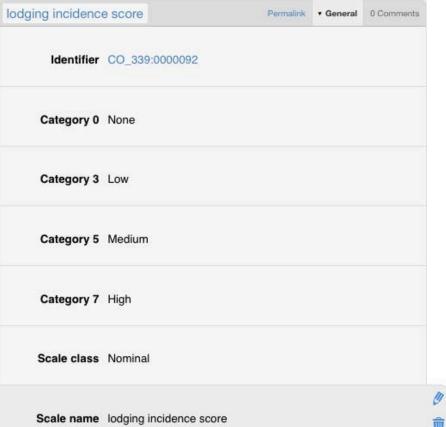


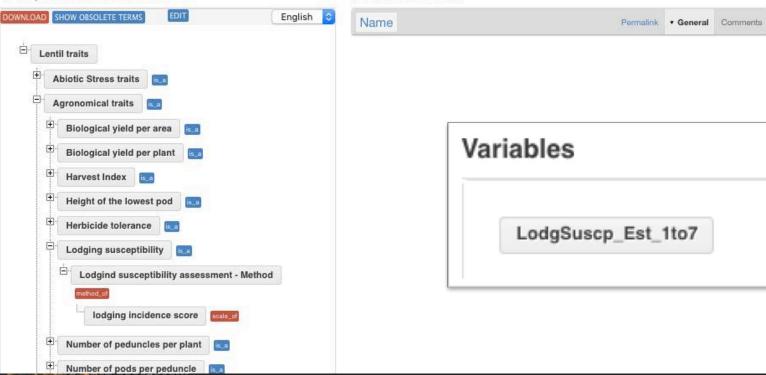
Logout

ElizabethArnaud

**Annotation Tool** 

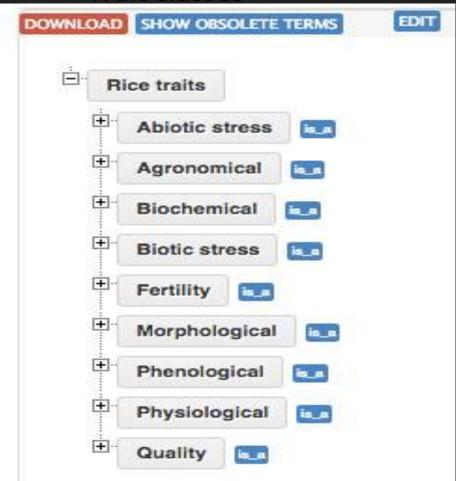








#### Trait classes



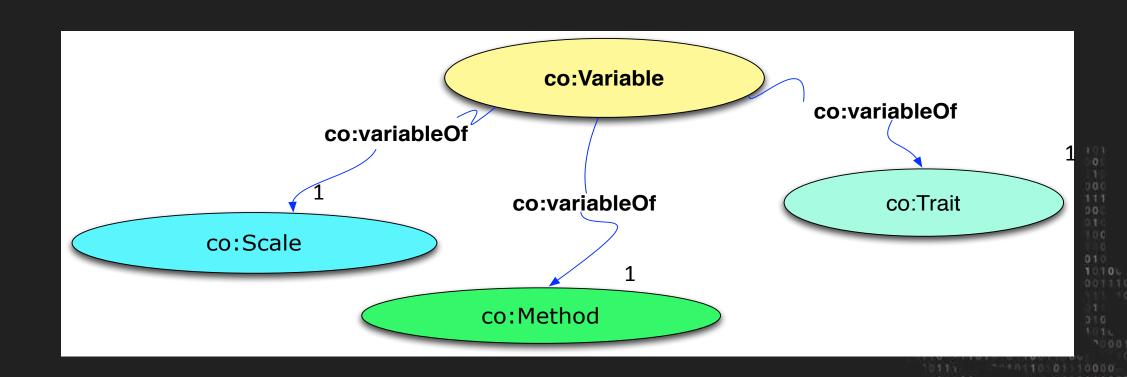
THE Methods, scales have a definition

Classific	ation of methods	Method class	Examples	
	By using a measuring device, a sensor	Measurement	Plant height measuring with a ruler     Fruit weighing on a scale	
The trait is	By counting entities	Counting	- Leaf counting	
i i	By an assessment that only relies on the experience and subjectivity of the observer. The assessment is not supported by a measuring device.	Estimation	<ul> <li>Grain colour estimation</li> <li>Damage on leaves visual estimation</li> <li>Plant height visual estimation</li> </ul>	
The trait observation derives from an aggregation of observations (regardless of how they were observed i.e., measured, counted, estimated or computed)		Computation	1000 grain weight calculation     (1000 * measured grain weight / grain count)      Harvest index calculation (Grain yield / Aboveground biomass)      Grain protein content calculation (Grain Nitrogen content * 6.25)	

#### **Classification of scales**

Classific	cation of scales
Method name	Description
Code	This scale class is exceptionally used to express complex traits. Code is a nominal scale that combines the expressions of the different traits composing the complex trait. For example a disease related code might be expressed by a 2 digit expressing intensity and 2 character code expressing the severity. The first 2 digits are the proportion of plants affected by a fungus and the 2 characters refer to the severity, e.g. "75HD" means "75% of the plants are infected and plants are highly damaged". It is recommended to create variables for every component of the code.
Duration	The date class is for time elapsed between two events expressed in a time format, e.g. "days", "hours", "months".
Nominal	Categorical scale that can take one of a limited number of categories. There is no intrinsic ordering to the categories.
Numerical	Numerical scales express the trait with real numbers. The numerical scale defines the unit e.g. centimeter, ton per hectar, branches.
Ordinal	Ordinal scales are composed of ordered and fixed number of categories.
Text	A free text is used to express the trait. Also known as character variable.
Date	The date class is for events expressed in a time format, e.g. "yyyymmdd hh:mm:ss – UTC" or "dd/mm/yy".

# **Example of an ontology**









### Improving the re-usability of the data file

The content of the Excel file is not always re-usable because of lack of clear variable name

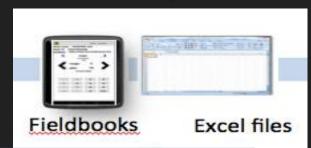
Germplasm ID	PH	PH2	GCOL	GY
24530	80	1	2	35
85432	120	3	3	48
78452	95	2	3	43
56093	100	2	1	50



Fieldbook design system







Germplasm ID	PltHei_Mesure_ cm	PltHei_Mesure_ category
24530	80	1
85432	120	3
78452	95	2
56093	100	2



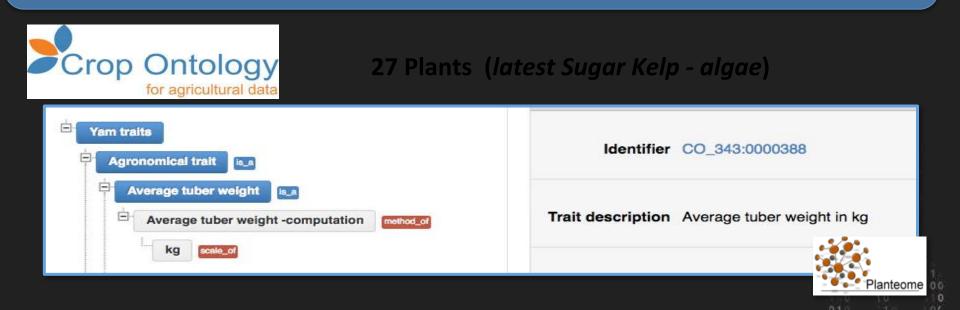
Data file using standard variables.



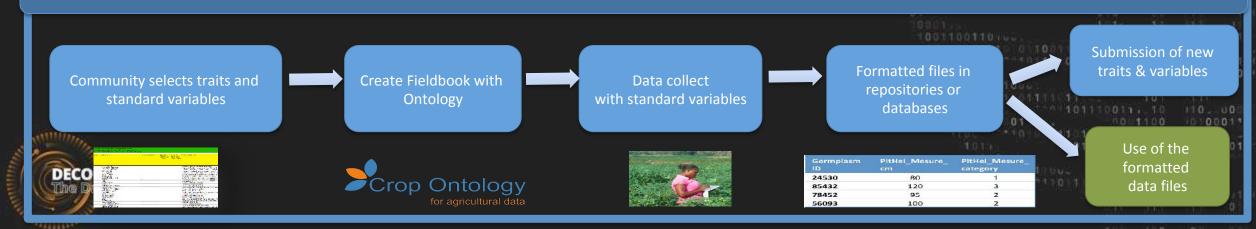


# Crop Ontology for Trait Data www.cropontology.org



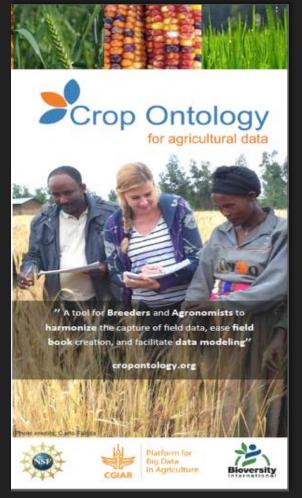








### **Crop Traits**

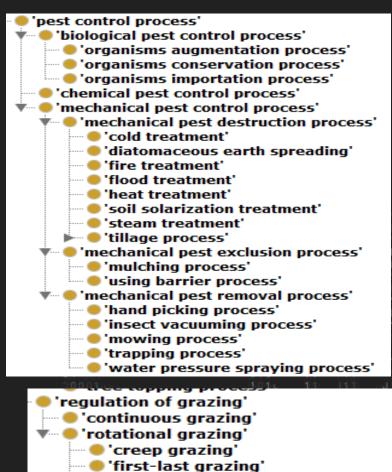


Provides names, definition and semantic relationships for measured traits and agronomic parameters with their standard variables for Agronomic and Breeding fieldbooks, surveys and for data annotation in the databases and repositories

Integrating Farmers preferences with respect to gender



### **Agronomy Ontology**



'intensive rotational grazing'

'hohenheim grazing system' 'holistic planned grazing'

'controlled grazing'

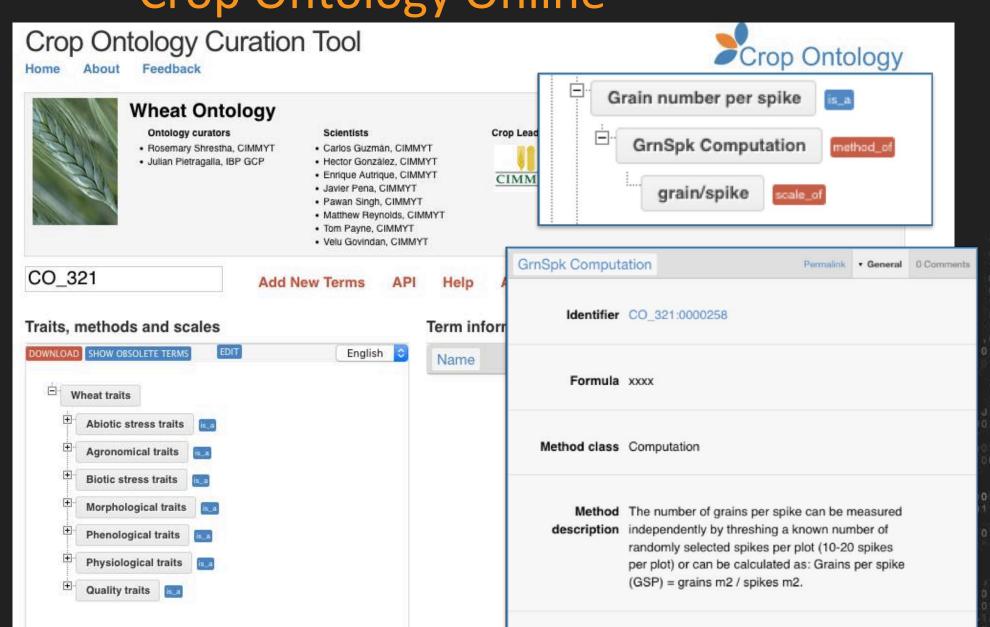
| 'mob grazing' | 'strip grazing'



# **Crop Ontology Online**

- 27 species
- Download in csv, obo,
   RDF
- API
- Helpdesk
- Adding farmers preferences, sensory traits



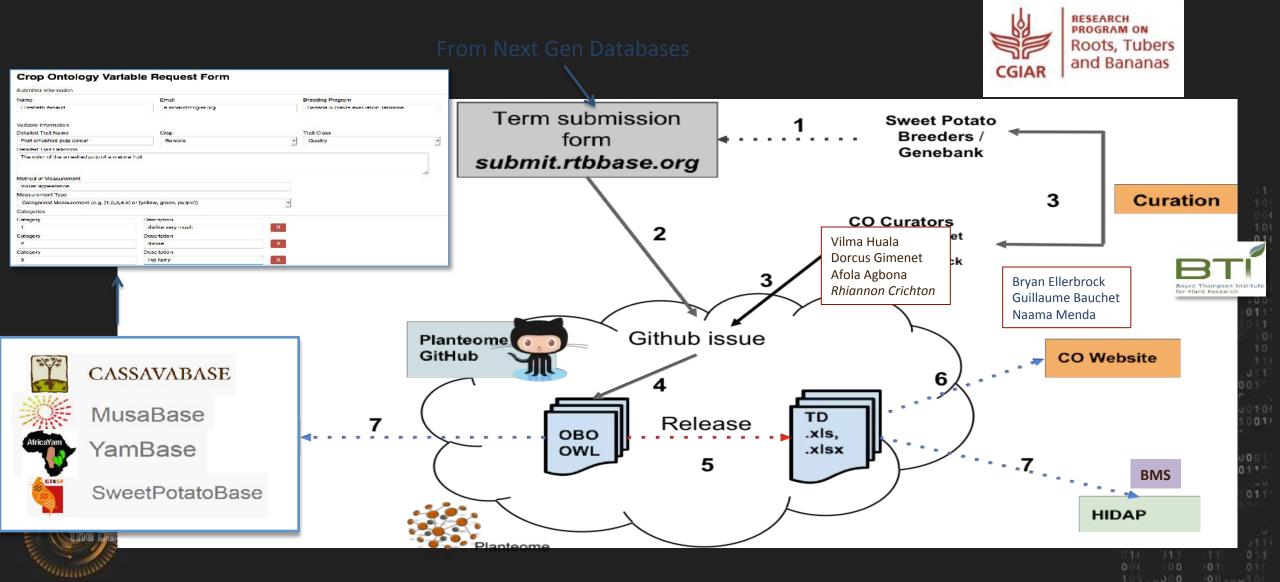


### Where is Crop Ontology applied?



Repository and Database

### **Crop Ontology Curation Workflow**





### Getting Involved: How can you contribute and utilize these resources?

- Explore annotations and ontologies for your own research work
- Create a GitHub account to request terms, make comments, or share your own ontology
- Annotate your data, and work with us to submit to our repository





# Linking the application Crop Ontology to the reference Plant Ontology of Planteome

From Species-specific to species-agnostic

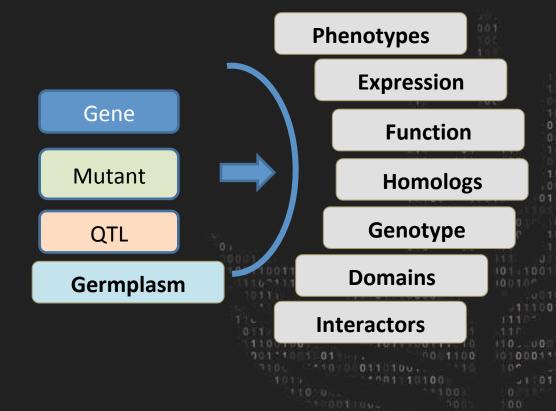




### Ontologies can link to data from multiple species

Examples of fruit color trait (TO:0002617) across various species







### **Species-Specific Ontologies and Data Annotations**

#### Cassava- IITA



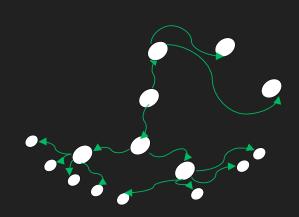
Image Source: Wikipedia

#### 163 traits, for example:

- stem number
- root neck length
- red spider mite severity

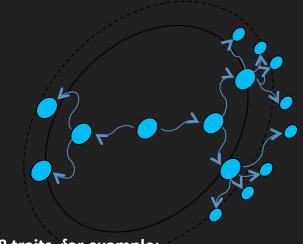
#### Lentil- ICARDA





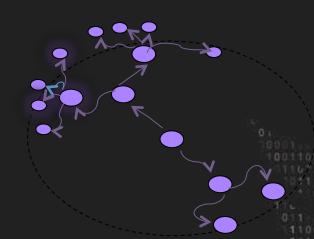
Three examples from CGIAR Centers integrated into Planteome:

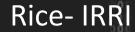
How can we make interspecific comparisons across these ontologies?



68 traits, for example:

- pod weight
- cooking time
- root length





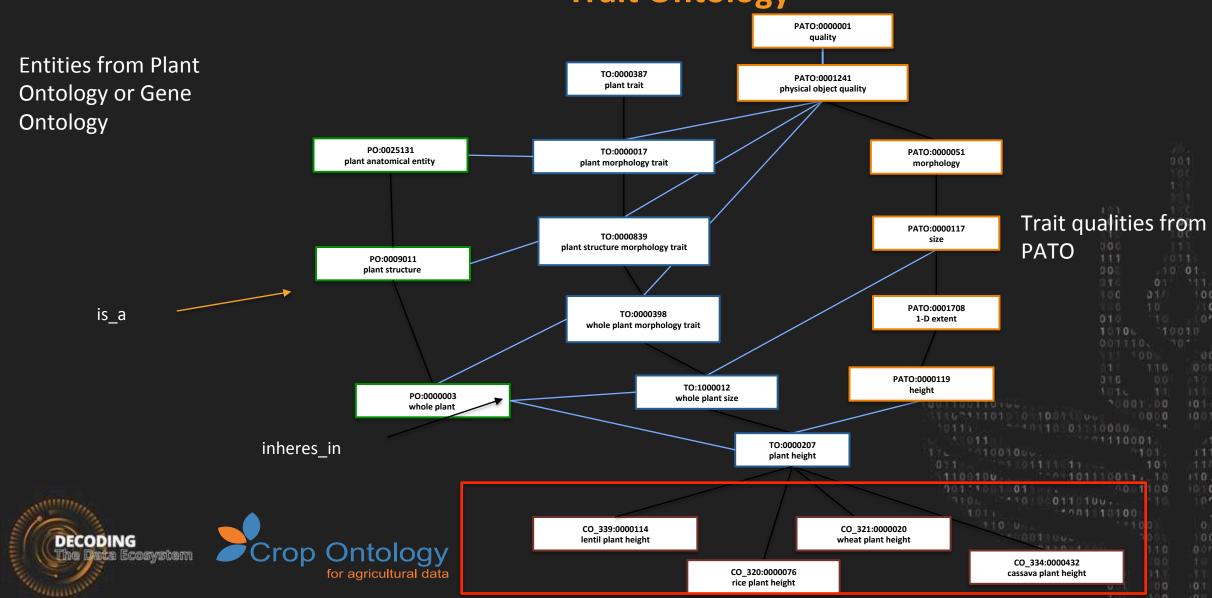


#### 157 traits, for example:

- caryopsis shape
- amylose content
- lodging incidence



# Integration of Species-Specific Ontologies in Planteome Reference Trait Ontology



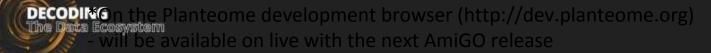


# Progress of CO to TO Mapping:

Planteome

Ten CO crop vocabularies have been mapped to the TO:

	# traits	# manually curated
CO_320_rice	157	5
CO_321_wheat	266	9
CO_322_maize	200	31
CO_324_sorghum	130	27
CO_331_sweetpotato	195	27
CO_334_cassava	163	16
CO_336_soybean	83	2
CO_339_lentil	68	11
CO_341_pigeonpea	62	9
*CO_343_yam	159	40



### **ACKNOWLEDGMENTS**

### Planteome Partners

#### **Oregon State University**

Pankaj Jaiswal (Lead PI), Laurel Cooper, Justin L. Elser, Austin Meier, Justin Preece



Eugene Zhang, Botong Qu



#### **Elizabeth Arnaud, Bioversity:**

Marie-Angélique Laporte, Leo Valette

#### CGIAR:

Rosemary Shrestha (CIMMYT), Karthika Rajendran (ICARDA), Agbona Afola (IITA), Omar Benites (CIP), Vilma Hualla (CIP), Jeffery Detras (IRRI) And many others!

Lawrence Berkeley National Lab

Christopher Mungall, Seth Carbon



NY Botanical Garden, USA

**Dennis Stevenson** 



University at Buffalo, NY Barry Smith



University of Birmingham, UK Georgios Gkoutos,





UC Santa Barbara

BS Manjunath (Bisque), Dimitry Fedorov, Kristian Kevilekval

#### **Collaborator Labs & Projects**

Users and resource databases who have adopted the Planteome Ontologies and contributed data:

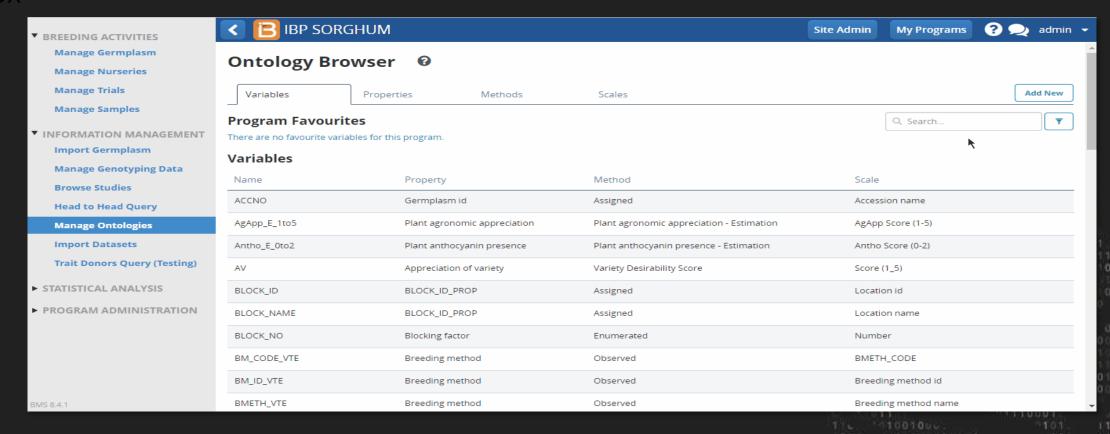
CyVerse, Gene Ontology Consortium, Protégé, OBO Foundry, Gramene, Ensembl Plants, USDA-GRIN, Rice Diversity Project, SolGenomics N, MaizeGDB, Rex Nelson, SoyBase, TAIR, Steve Cannon, LegumeIS DivSeek

And many others!

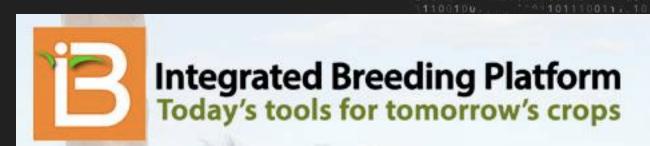


### Ontology Manager of the Breeding Management System

UX









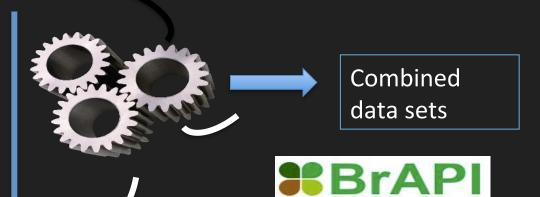
# CO variable format used by MIAPPE & Standard Breeding API (BrAPI)

Germplasm Genotypes

Multi Crop Passport Data







Breeding Management System



The Part VICAPPEN = Minimum Information About a Plant Phenotyping Experiment

# Harmonizing Agronomic Data





# Agronomy Ontology and Fieldbook



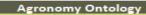


#### A collaborative project

A Community of Practice of agronomists and data managers is coalescing to contribute to the ontology content, test the agronomy fieldbook within ongoing projects, and provide feedback. Scientists from CIRAD, INRA, Crops for the Future, NARO, CIMMYT, have already showed their interest in participating.



The Data Ecosystem



Describes variables commonly used in agronomic experiments

- 550 agronomic classes and 230 variables
- Gathered ICASA and CIMMYT1 variables
- Built in the Ontology Web Language format with Protégé

Web application that lists variables of the **Agronomy Ontology** to ease the creation

of fieldbook

Data collection and harmonization tool by and for the community Agronomy gronomy Fieldbook

Annotated data

Store data annotated with variables defined in the Agronomy Ontology



Researcher

Designs data collection

sheets on field book web

application

Data collection Annotated data



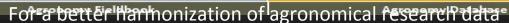
Data collector Collects data using data collection sheets (Excel sheet or mobile app)















## Development of the Agronomy Ontology (AgrO)

#### **Trial Information**

Person
Experimental design
Site informations
Institute informations

### Agronomy Ontology

Socio-Economy

#### Agricultural operations

Land preparation
Seeding
Fertilizer
Irrigation
Pest control
Harvest
Agricultural equipment

#### Monitored data

#### Abiotic factors

Soil
Rainfall
Light
Temperature
Humidity
Atmosphere

**Biotic factors** 

**Crop Ontology** 

**Pest Ontology** 

Germplasm

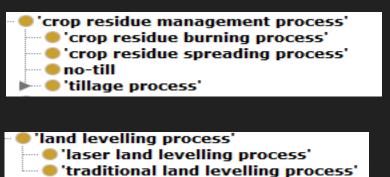


### AgrO Content

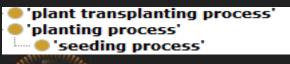




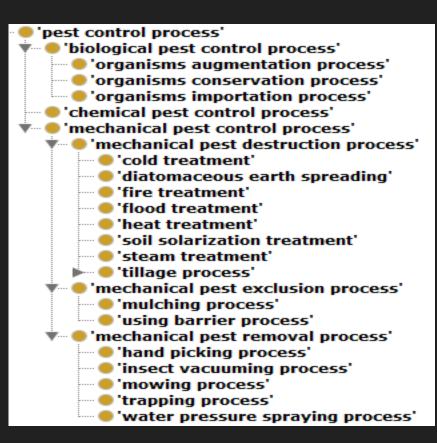
agricultural implements, inputs, agricultural practices Entity and a Quality (e.g.: field (entity) area (quality))

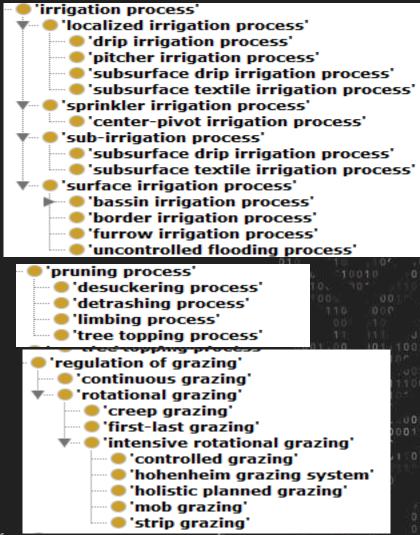




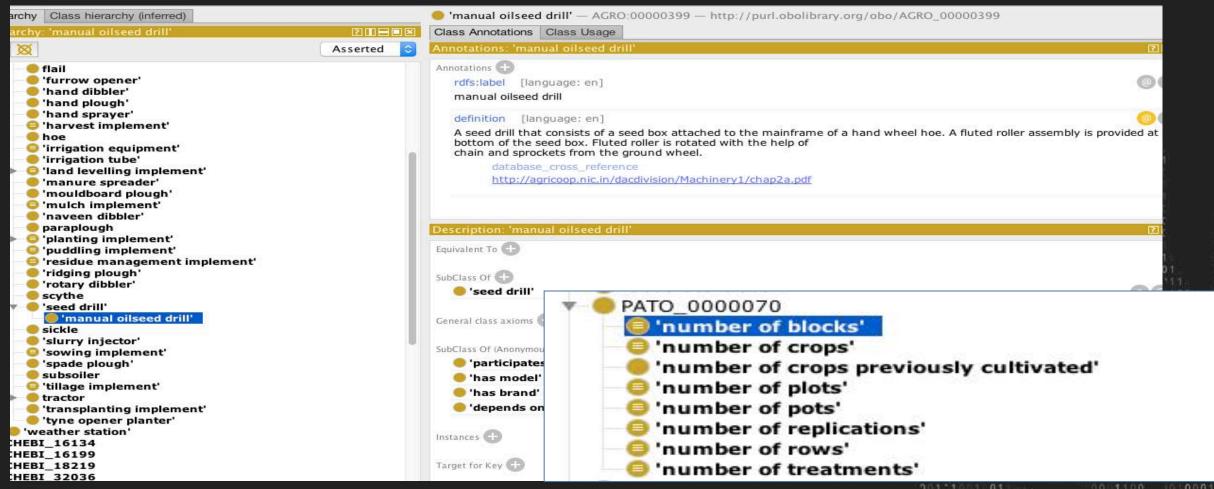








## AgrO - a composite ontology in OWL





Quality classes are imported from PATO (Phenotypic Quality Ontology), whereas the unit classes are imported from the Unit Ontology

#### AgroFIMS – user-tested; ready to field-test in Spring 2019







Hello, Guest O Not connected

**◆**0 Log in



III Single Trial Analysis

2 Documentation

Help

About

### HIDAP AgroFIMS

#### Agronomy Field Management System

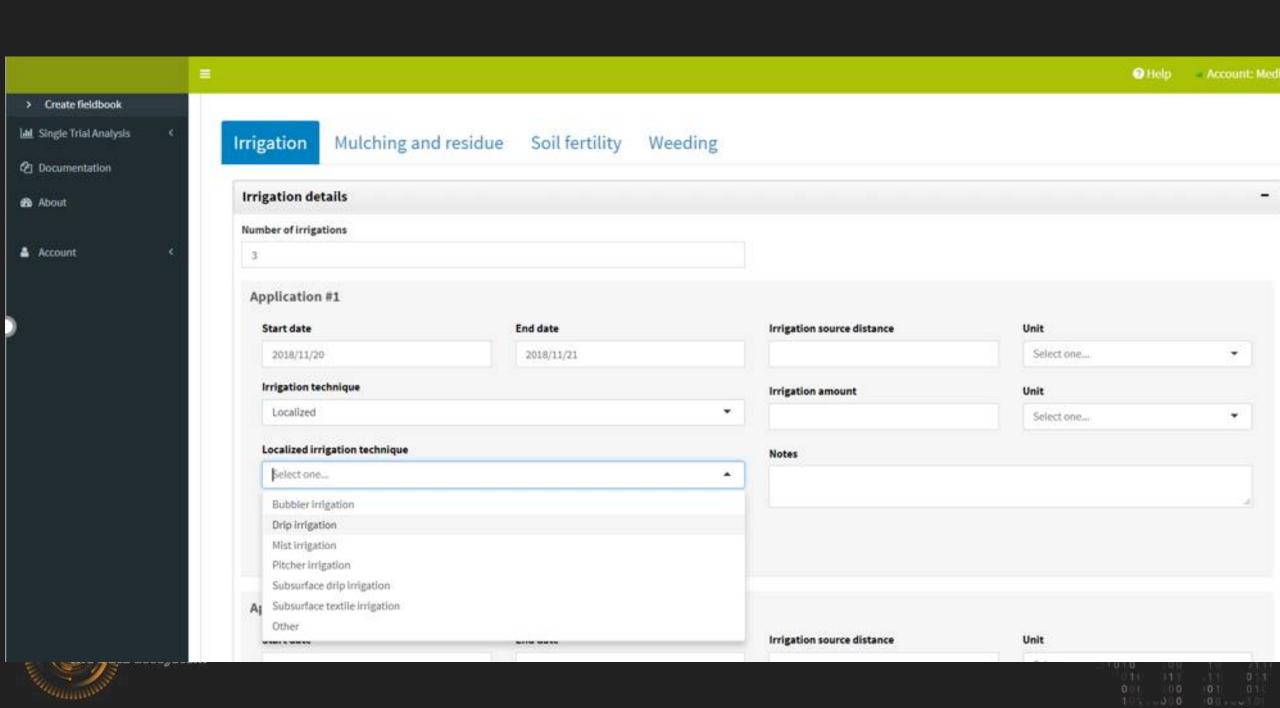


#### HIDAP AgroFIMS v0.0.17

The Agronomy Field Information Management System (AgroFIMS) has been developed on CGIAR's HiDAP (Highly-interactive Data Analysis Platform created by CGIAR's International Potato Center, CIP). AgroFIMS draws fully on ontologies, particularly the Agronomy Ontology and the Crop Ontology. It consists of modules that represent the typical cycle of operations in agronomic trial management, and enables the creation of data collection sheets using the same ontology-based set of variables, terminology, units and protocols. AgroFIMS therefore:

- Standardizes data collection and description for easy aggregation and inter-linking across disparate datasets;
- · Allows easy integration with HiDAP breeding data, or any other ontology-based datasets;
- · Functions as a data staging repository, allowing data uploads with view/edit permissions;
- · Enables data quality checks, statistical analysis of the data collected, and the generation of sophisticated statistics reports;
- · Aligns a priori with CGIAR's CG Core metadata schema;
- · Enables easy upload to the institutional repositories, and much more.

Funding for AgroFIMS was provided by the Bill and Melinda Gates Foundation's Open Access, Open Data Initiative, and the CGIAR Big Data Platform.



#### Crop measurement

Please, select measurement by click.

Show 25	▼ entries			Select all	Search:	
	Crop	Group	Subgroup	Crop measurement	Scale	\$
1	Cassava	General	Timing	Date	yyyy/mm/dd	
2	Cassava	Biomass	Harvest	Area harvested	m2	
3	Cassava	Biomass	Storage roots	Fresh weight	g	
4	Cassava	Biomass	Storage roots	Subsample fresh weight	g	
5	Cassava	Biomass	Storage roots	Subsample dry weight	g	
6	Cassava	Biomass	Storage roots	Moisture content	%	
7	Cassava	Biomass	Storage roots	Dry weight	g	
8	Cassava	Biomass	Storage roots	Dry matter yield	kg/ha	
9	Cassava	Biomass	Leaves	Fresh weight	g	
10	Cassava	Biomass	Leaves	Subsample fresh weight	g	
11	Cassava	Biomass	Leaves	Subsample dry weight	g	
12	Cassava	Biomass	Leaves	Moisture content	%	
13	Cassava	Biomass	Leaves	Dry weight	g	
14	Cassava	Biomass	Leaves	Dry matter yield	kg/ha	
15	Cassava	Biomass	Stems	Fresh weight	g	



## ICRISAT Dataverse: Data set annotation with ontologies Annotation on Dataverse

#### to 10 of 254 Results

IT Sort .

Phenotypic evaluation data of Pigeonpea (C.acutifolious) medium duration advanced lines trial for year 2016-17



Nov 14, 2018 - Phenotypic

Sameer Kumar, CV; Anupama Hingane, 2017, "Phenotypic evaluation data of Pigeonpea (C.acutifolious) medium duration advanced lines trial for year 2016-17" doi:10.21421/D2/XALIZED ICRISAT Datayerse V3. LINE:6:Elf26uVyetmS2hWMy/QeAWQ--

#### Keyword

Chickpea (CROP ONTOLOGY) http://www.cropontology.org/terms/CO 338:ROOT/

Plant height (CROP ONTOLOGY) http://www.cropontology.org/terms/CO\_338:0000270/

Days to maturity (CROP ONTOLOGY) http://www.cropontology.org/terms/CO\_338:0000281/

Days to 50% flowering (CROP ONTOLOGY) http://www.cropontology.org/terms/CO 338:0000279/

Replication number (CHOP UNIOLOGY) http://www.cropontology.org/terms/CO 715:0000246/

Entry number (CROP ONTOLOGY) http://www.cropontology.org/terms/CO\_715:0000031/

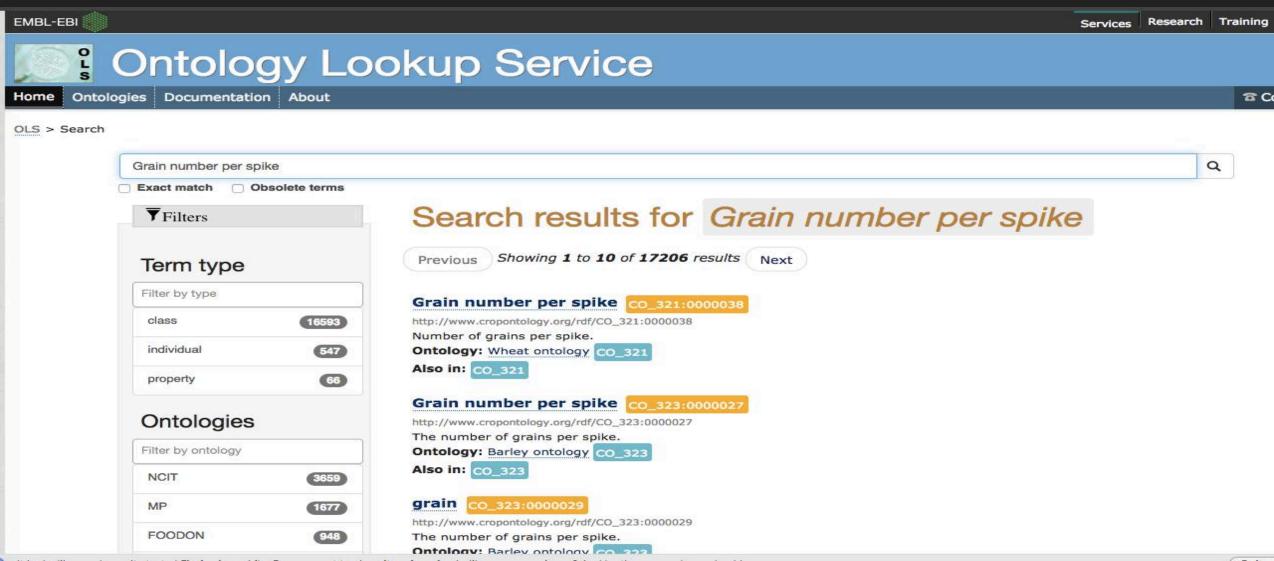
Seed yield per plot (CROP ONTOLOGY) http://www.cropontology.org/terms/CO 338:0000271/

Cood viold (CDOD ONTO) OCV http://www.orongotology.org/torms/CO 220:0000275/												
A	В	C	D	Ē	F	G	H		J	K		
Plot No	Replication Number	Entry No	Entry name	Days to 50 percent flowering	Days to maturity	Plant height	Plant count at harvest	Seed yield per plot	Seed yield	Hundred seed weigh		
401	1	1	ICCV 14301	47	96	39,4	90	457,33	953	35,1		
402	1	8	ICCV 14308	47	97	38,4	70	570,13	1188	40		
403	1	20	Local Check	48	98	46,6	75	696,00	1450	39		
404	1	16	ICCV 14316	47	98	41	75	738,76	1539	45,5		
405	1	19	KAK 2	46	96	45,8	102	618,00	1287	38		
406	1	14	ICCV 14314	46	95	44,8	91	959,00	1998	39,5		



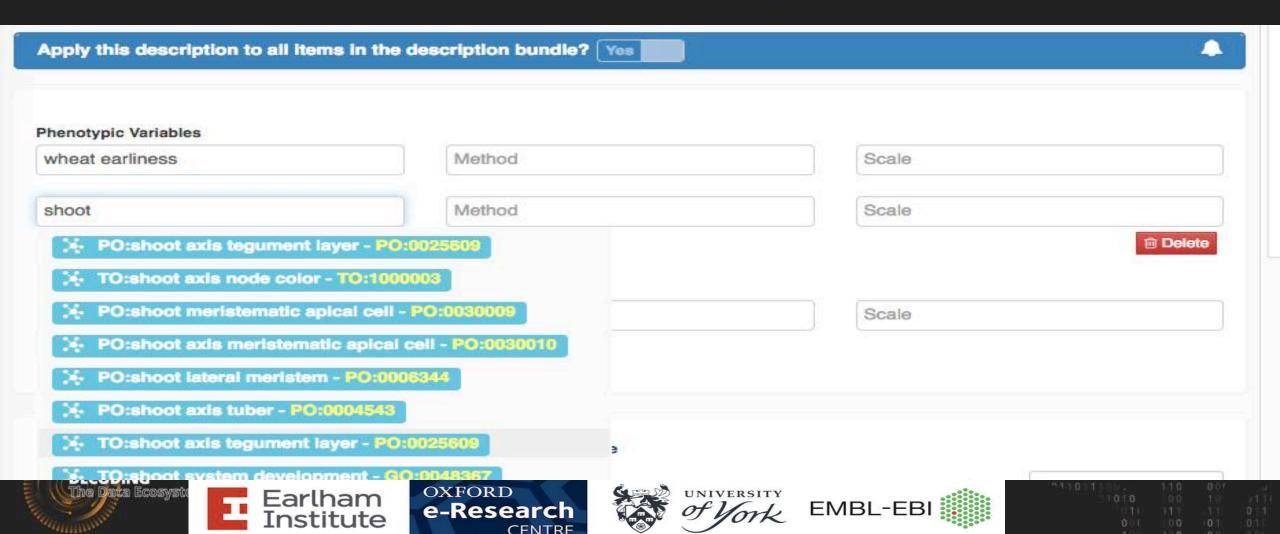


## Ontology Look Up Service European Bioinformatics Institute





# Data Annotation tool for public repositories using the OLS <a href="https://copo-project.org/">https://copo-project.org/</a>



### Multi-disciplinary data for agriculture research questions

- In Nigeria, Rainfed lowlands = 70% of total rice area.
- Yield losses resulting from flooding may range from 10% to total crop loss.

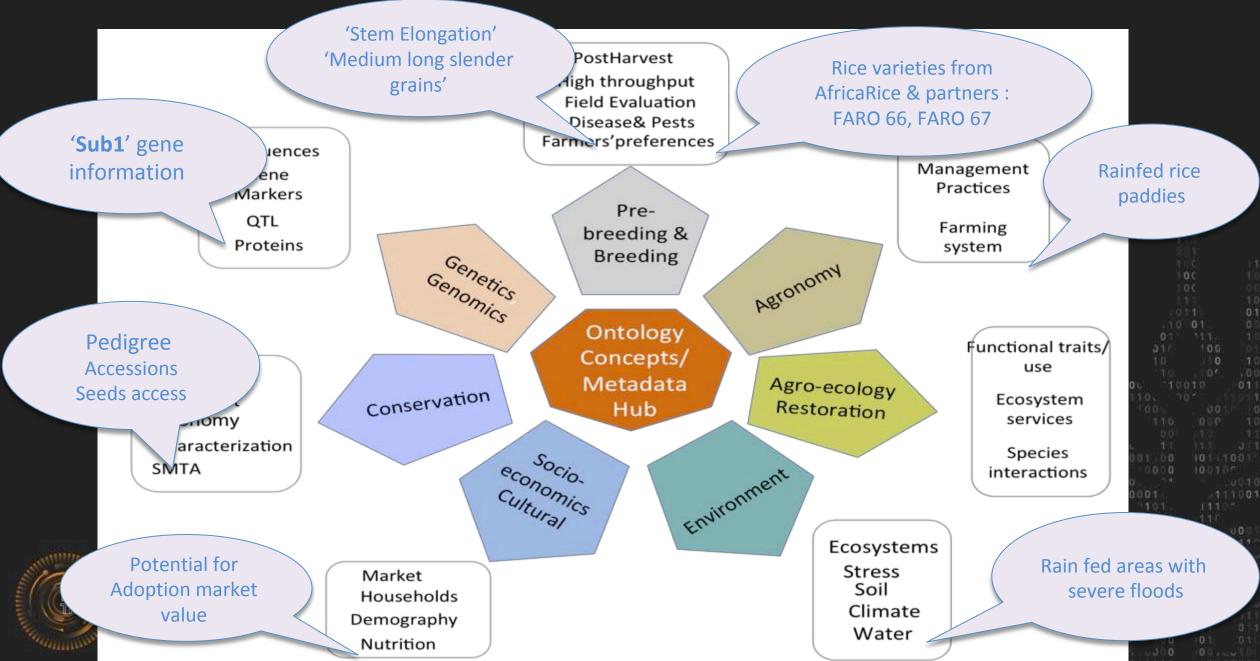
#### Question:

Identify rice <u>varieties</u> that are <u>high-yielding and</u> <u>flood-tolerant</u> that can grow in <u>Nigeria</u> in <u>rain-fed lowlands subject to recurrent devastating flooding?</u>





### Multi-disciplinary Agricultural Data





### Multi-disciplinary Agricultural Data

**PostHarvest** 

Ontology

Sequence Ontology Gene Ontology

**MCPD Taxonomy** Ontology Descriptors

> Germplasm **Passport Taxonomy** Characterization

Gene

Markers

**Proteins** 

QTL

**Social Ontology** SocioecO\* ANT\*

DECO

High throughput **Field Evaluation** Disease& Pests Farmers'preferences Seguences

> Prebreeding & Genetics Breeding Genomics

Concepts/ Metadata Conservation Hub

> Socio. economics Cultural

Market Households Demography Nutrition

Crop Ontology\* **Animal Trait Ontology** Plant Diseases

Agronomy

Environment

Agro-ecology

Restoration

**Ecosystems** 

Climate

Water

Stress

Soil

Management **Practices** 

> **Farming** system

> > Functional tra use

> > > Ecosystem services

**Species** interactions Agronomy Ontology\* **AGROVOC** 

> **ECOCORE Plant Trait** Thesaurus **SDGiOntology**

**Environment** Ontology



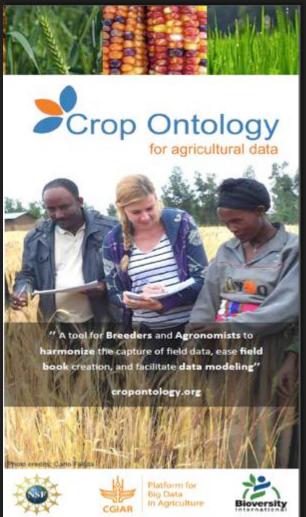
## **Ontology Community of Pratice**







#### **Crop Traits**







### **Agronomy Ontology**

pest control process' biological pest control process' 'organisms augmentation process' 'organisms conservation process' 'organisms importation process' 'chemical pest control process' 'mechanical pest control process' 'mechanical pest destruction process' 'cold treatment' 'diatomaceous earth spreading' 'fire treatment' 'flood treatment' 'heat treatment' 'soil solarization treatment' 'steam treatment' 🛌 🔴 'tillage process' mechanical pest exclusion process' 'mulching process' 'using barrier process' 'mechanical pest removal process' 'hand picking process' 'insect vacuuming process' 'mowing process' 'trapping process' 'water pressure spraying process' 'regulation of grazing' 'continuous grazing' 'rotational grazing' 'creep grazing' 'first-last grazing' 'intensive rotational grazing' 'controlled grazing' 'hohenheim grazing system' 'holistic planned grazing' 'mob grazing'

'strip grazing'



#### SOCIO-ECONOMIC ONTOLOGY (Socio)

A socio-economic ontology of controlled vocabularies, classifications, and concordances that allow standardization of key indicators. The key is that in SociOl is a 'living' ontology, since the high variety of socio-economic data implies that it will need to evolve continuously.











Joint effort from two Community of Practices (CoP)

The SociO ontology has been developing from the collaborative work between the ontologes Data CoP and socio-economic data CoP in the CGIAR the Platform for Big data in Agriculture.



The coverage of key socio-economic indicators

The SociO ontology team work together with the <u>working group on 100</u> guestions in standard household surveys in order to get information on key socio-economic indicators that can be useful for research, policymaking, and impact assessment.



Linking existing ontologies

The SociO ontology team is working with external partners to re-use existing ontologies such as Basic Formal ontology, Information Artifact ontology, Informed Content Ontology, CGIAR agronomy ontology and others, so that we can easily link to other ontologies in the Linked Open Data cloud.



SociO ontology for annotating survey data

The SociO ontology would be used for annotating socio-economic survey data from a survey design stage (i.e. questionnaires) to data sharing stage in the data management life cycle.

Collaboration of the Ontology CoP and the Socio Economic Data CoF

For household surveys





## Discussion and Working Groups



**Livestock Ontology** 

Plant Phenotype

**Genetics and Genomics** 

**Plant Stress** 



Fish Ontology



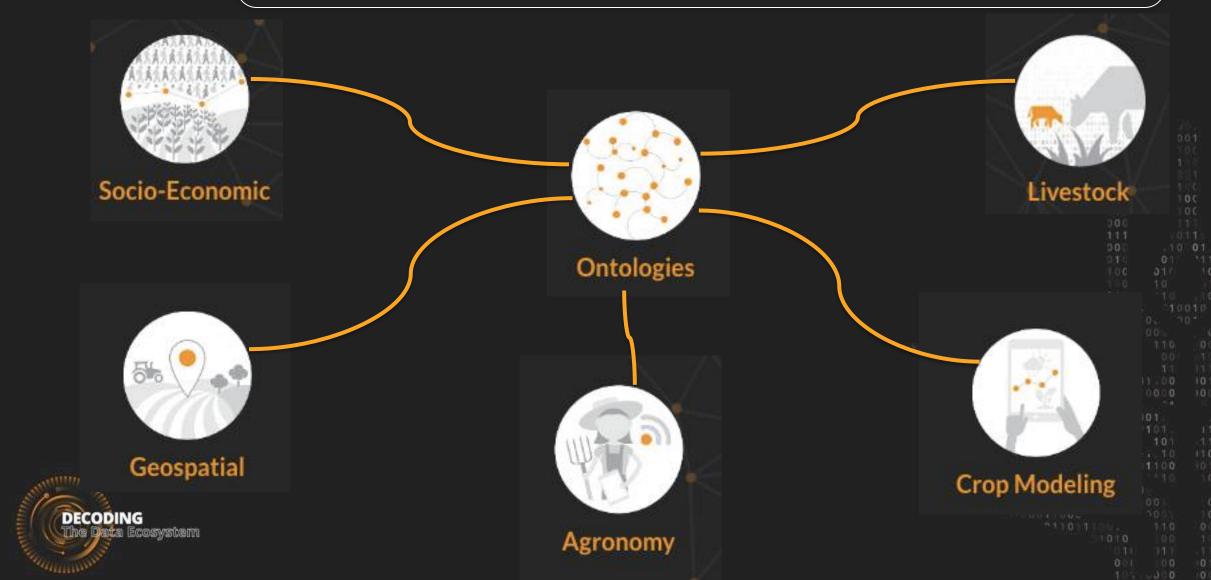
Water Management







Adopt, adapt and validate best practices, guidelines to support the harmonization and interoperability of data needed by the Big Data Platform and to support the other Communities of Practice.





### **A Growing Community Of Practice**

2018









# Interested in joining our community of practice?

Sign up to our mailing list for community news and updates.

VS



and for data interoperability and data discovery following the FAIR principles.

This space can be used as a discussion area, share and