Developing and sharing ontologies: a key step towards efficient genetic and breeding strategies, a sorghum case study

David Pot : david.pot@cirad.fr

cirad LA RECHERCHE AGRONOMIQUE POUR LE DÉVELOPPEMENT

agap

AIDA Unité de recherche Agroécologie et intensification durable des cultures annuelles

iavao Innovation et amélioration variétale en Afrique de l'Ouest

Crop Ontology for agricultural data

Talk's Outline

My expectations, to reach my objectives, in relation with the description of phenotypic data ?

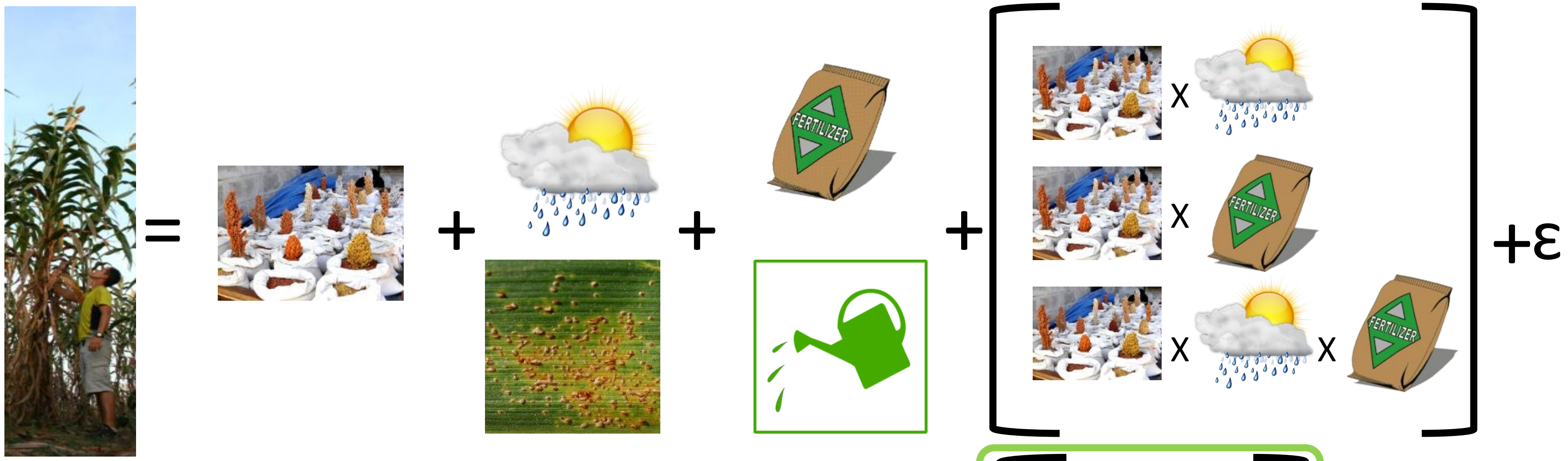
What do I need to efficiently describe traits and make them useful for me and the communities ?

How to optimize the development of controlled vocabularies and extend their uses ?

What are my expectations, in relation to phenotypic data, to reach my objectives?

As a geneticist involved in a breeding programme, my objectives are:

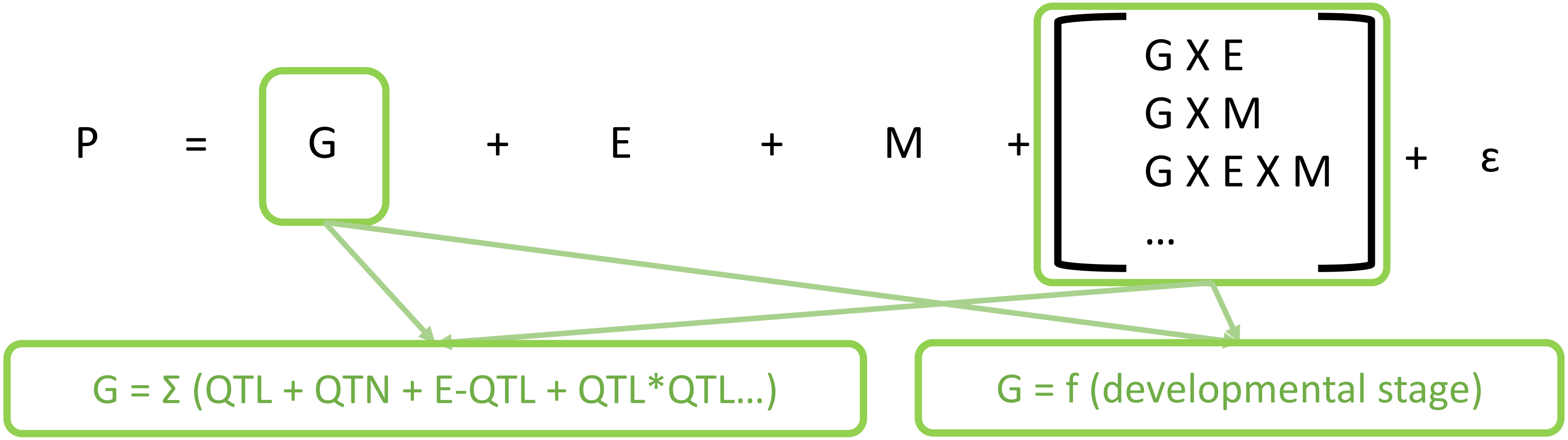
- To develop varieties (growers and end-users)
- To predict their phenotype(s) in different contexts



$$P = \boxed{G} + E + M + \boxed{\begin{bmatrix} G \, X \, E \\ G \, X \, M \\ G \, X \, E \, X \, M \\ \ldots \end{bmatrix}} + \varepsilon$$

- **Estimating G…(on available trials) and predicting P (in non tested environments for tested and non tested genotypes)**

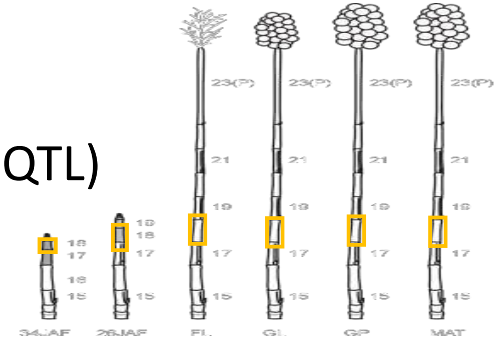
Estimating G and GxN to predict P,
Dissecting G components



P = G + E + M + [G X E, G X M, G X E X M, …] + ε

G = Σ (QTL + QTN + E-QTL + QTL*QTL…)

G = f (developmental stage)



- **I want to:**
  - Identify genomic locations (QTL, QTN, e-QTL)
  - Highlight gene expression networks
  - Estimate breeding values
  - Predict phenotypes

- **I need take to advantage of multi-environment trials and datasets from others**

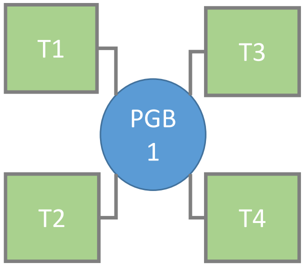
Improving estimation and prediction accuracies of G, GXE, GXEXM

Requires aggregating trials and studies

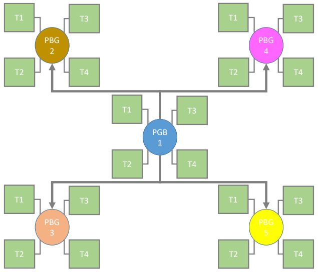
$$P = \boxed{G} + E + M + \boxed{\begin{bmatrix} G \times E \\ G \times M \\ G \times E \times M \\ \ldots \end{bmatrix}} + \varepsilon$$

- 1 trial : merging G with E and M, not efficient to estimate G
- 2 trials : « starting » to better estimate G effect
- n trials : much more confidence on G and G X n estimates

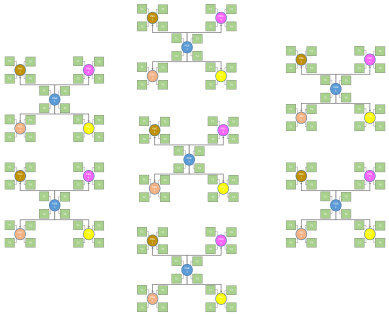
Me



My Network



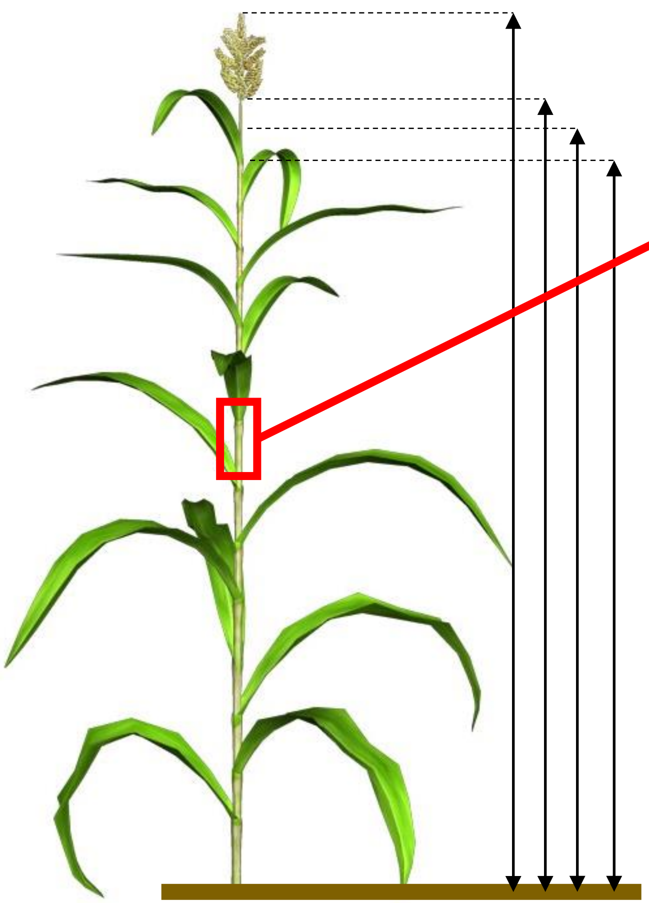
My species



- **I need to have accurate « machine readable » trait descriptions**
  => controlled vocabularies

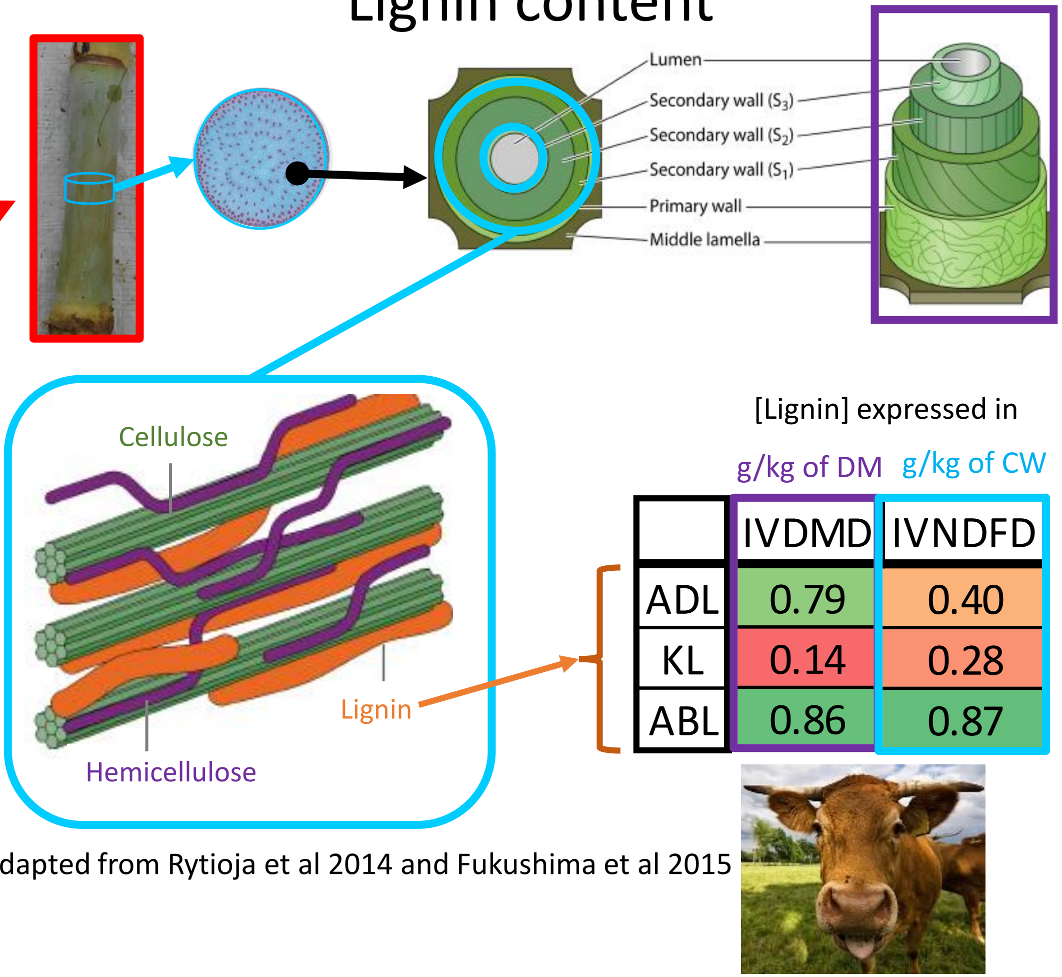
Trait description accuracy:
Current publications

Plant Height



=> **One trait = 4 traits…**

Lignin content



Adapted from Rytioja et al 2014 and Fukushima et al 2015

[Lignin] expressed in

	g/kg of DM	g/kg of CW
	IVDMD	IVNDFD
ADL	0.79	0.40
KL	0.14	0.28
ABL	0.86	0.87

=> **One trait = 6 traits…**

Trait description accuracy:
Identifying genomic regions of interest (QTL/ QTN)

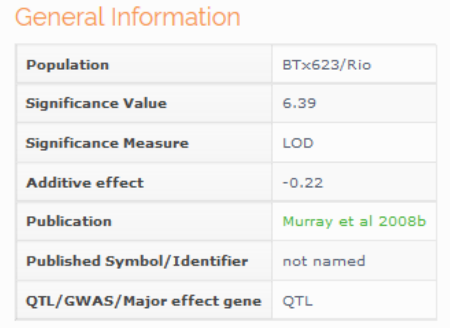
Sorghum QTL Atlas

Home > Sorghum QTL Atlas

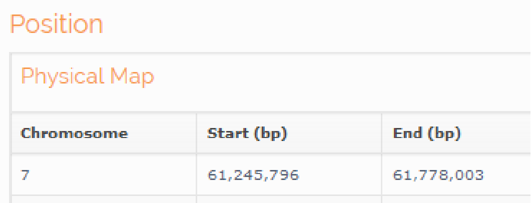
https://aussorgm.org.au/sorghum-qtl-atlas/ Mace, 2018

- 150 “QTL and GWAS” studies
- 191 traits
- ≈6000 QTL / QTN

- Lignin content
  - Trait Subcategory : "Stem composition"
  - Trait Description : "Lignin «
  - =>8 genomic regions (QTL / QTN)

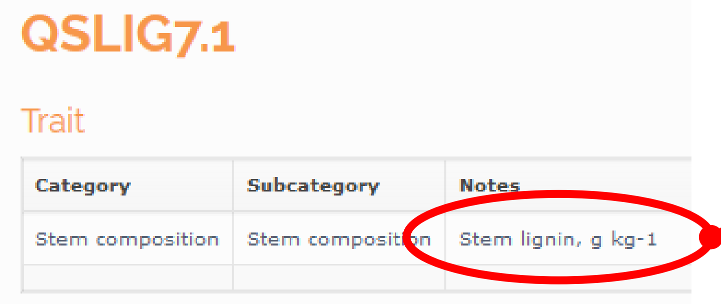

General Information

Population	BTx623/Rio
Significance Value	6.39
Significance Measure	LOD
Additive effect	-0.22
Publication	Murray et al 2008b
Published Symbol/Identifier	not named
QTL/GWAS/Major effect gene	QTL


Position

Physical Map

Chromosome	Start (bp)	End (bp)
7	61,245,796	61,778,003


QSLIG7.1

Trait

Category	Subcategory	Notes
Stem composition	Stem composition	Stem lignin, g kg-1

Method? Not clear in the paper

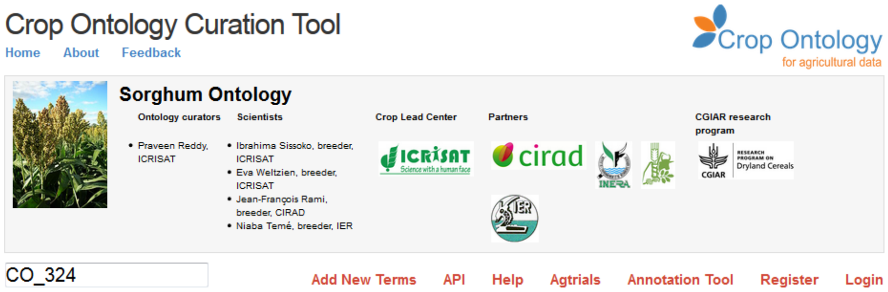
Ontogenic stage? Dough grain stage?

Unit? Dry Matter of stem?

- Need a more accurate definition of the traits (machine readable)

What do I need to efficiently describe my traits and make them useful for me and the community ?

The Crop ontology a relevant framework
to describe « Variables »



Crop Ontology Curation Tool

Home About Feedback

Crop Ontology for agricultural data

Sorghum Ontology

Ontology curators
- Praveen Reddy, ICRISAT

Scientists
- Ibrahima Sissoko, breeder, ICRISAT
- Eva Weltzien, breeder, ICRISAT
- Jean-François Rami, breeder, CIRAD
- Niaba Temé, breeder, IER

Crop Lead Center

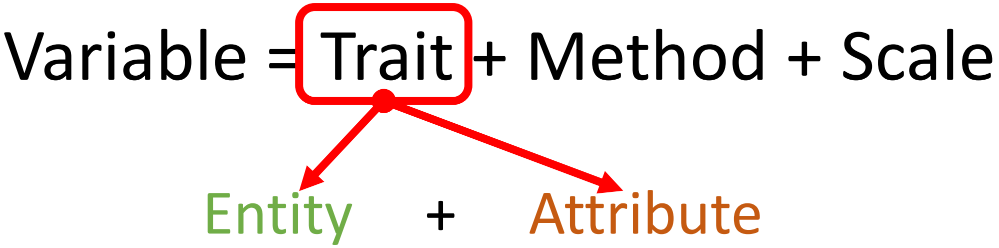
Partners

CGIAR research program

CO_324

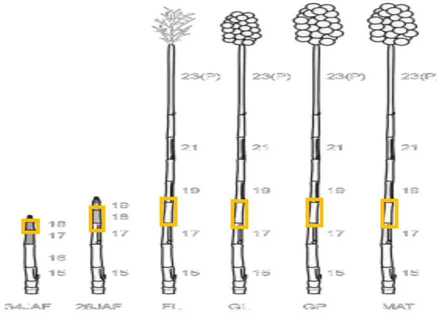
Add New Terms API Help Agtrials Annotation Tool Register Login

http://www.cropontology.org/ontology/CO_324/Sorghum



Variable = Trait + Method + Scale

Entity + Attribute



Organ:
- Internode
- Stem
- Whole aboveground biomass
- Group of cells…

What we measure
- Length
- Height
- Weight
- Area…

The Crop ontology a relevant framework

That needs to take advantage of others Ontologies

Plant trait ontology (TO):

- Lignin
- 4 traits
- TO:0000731

Lignin content (ABL) of IN18 120°C after elongation start in mg/g of cell wall residues

Plant Anatomy ontology (PO):

- Internode
- 34 organs
- PO:0020142/PO:0005005

Plant Anatomy ontology (PO):

- Internode development stages
  - 18 « stages »
  - Number of IN (max=16)
  - Elongation start
  - Elongation stage
- No information on level
- No information on age

Units Ontology (UO):

- Milligram per g
- Not found
- UO:0000308 : mg per kg
- kg of what ?

Environmental Ontology (ENVO):

- air temperature
- ENVO_09200001

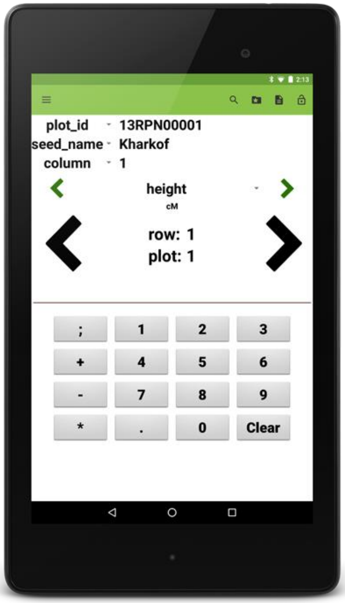
- **Method: no Ontology (?)**
  - ABL vs KL vs ADL
  - Prediction vs reference values (where ?)
  - Precision about unit (reference in proportion / ratio ?)

From the « CO » variables
to the fields, databases, papers, and to aggregations

Crop Ontology
for agricultural data

CO_324:0000623 :
PH_M_cm

Field book
alias
« Height »



plot_id 13RPN00001
seed_name Kharkof
column 1

height
cM

row: 1
plot: 1

http://wheatgenetics.org/fieldbook

« Local » DB

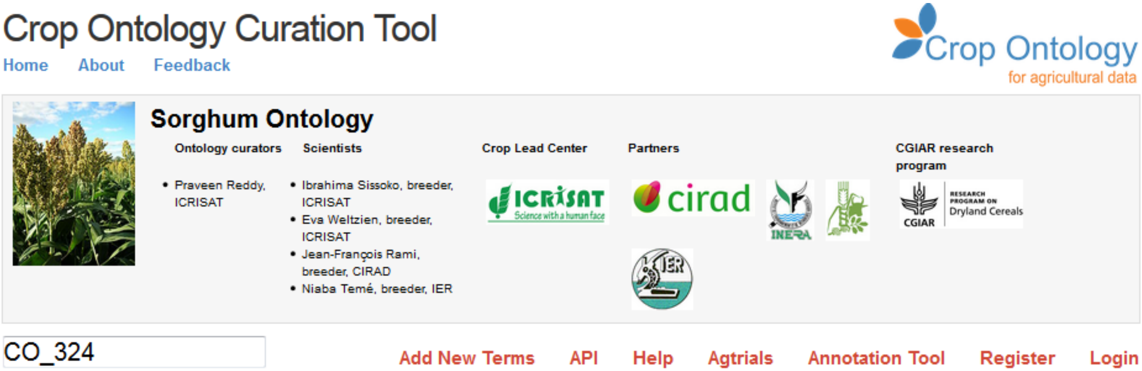
Daphne / AEGIS
BMS
PHISE
…

Publication

Open data

- **Data re-use**
- **Power and accuracy increases**

We have the framework,
We miss the « useful » content



Crop Ontology Curation Tool

Home About Feedback

Crop Ontology for agricultural data

Sorghum Ontology

Ontology curators
- Praveen Reddy, ICRISAT

Scientists
- Ibrahima Sissoko, breeder, ICRISAT
- Eva Weltzien, breeder, ICRISAT
- Jean-François Rami, breeder, CIRAD
- Niaba Temé, breeder, IER

Crop Lead Center

Partners

CGIAR research program

CO_324

Add New Terms API Help Agtrials Annotation Tool Register Login

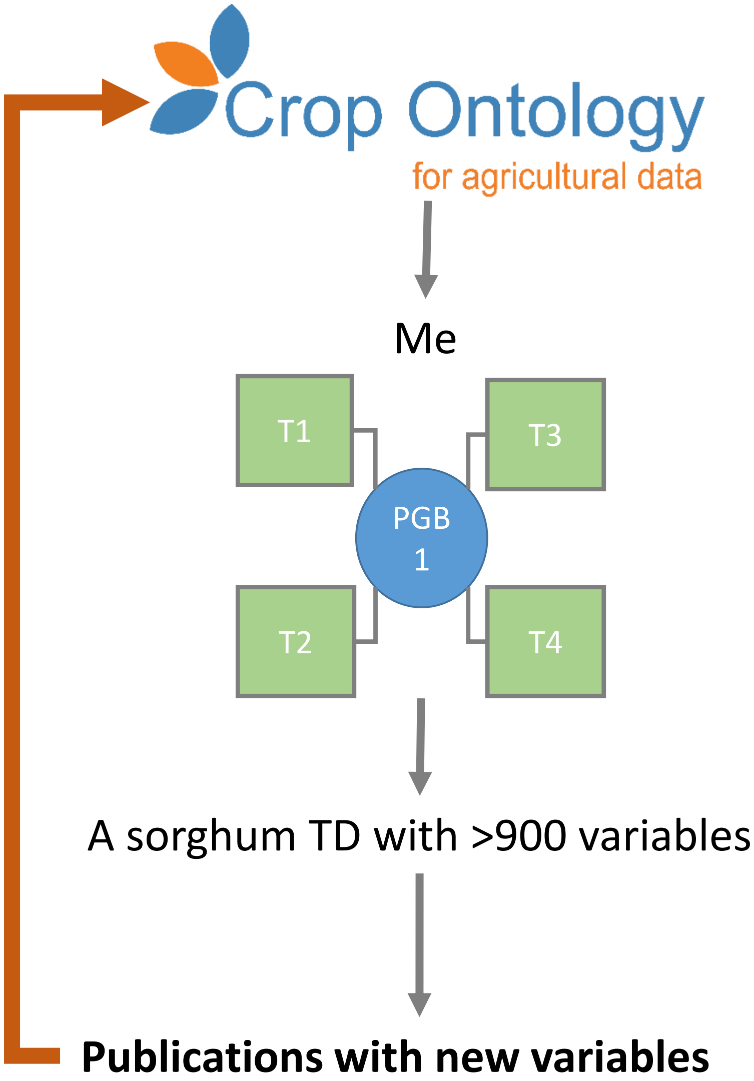
ICRISAT Science with a human face
cirad
INERA
IER
CGIAR
RESEARCH PROGRAM ON Dryland Cereals

- Sorghum TDv5 - Jan 2018
- 174 variables

- 2017-2018 : >50 articles « Sorghum » « Genetics »
  - 1 article (Guitton et al 2018) with trait ontology information (4 traits)…
- Before 2017 : no paper with reference to CO_324
- **We need to improve the referencing of the published traits**

How to optimize the development of controlled vocabularies: 3 propositions

Feeding the Crop ontology
« Individual inputs »



Crop Ontology

for agricultural data

Me

T1

T3

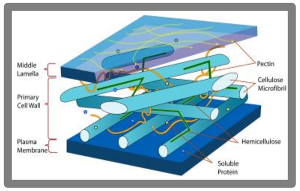
PGB 1

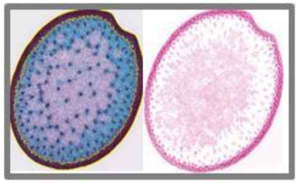
T2

T4

A sorghum TD with >900 variables

Publications with new variables

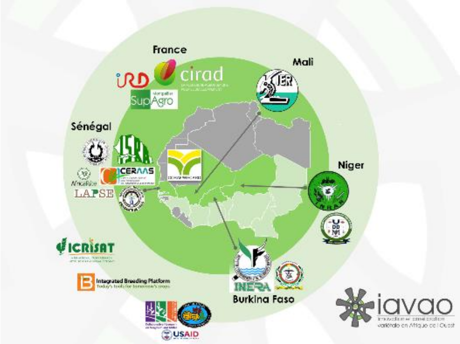




Feeding the Crop ontology

Direct contributions of Communities of experts

iavao
Innovation et amélioration
variétale en Afrique de l'Ouest



- Setting up Sorghum multi-environments trials in West Africa (basic trials)
- 9 Breeders from 5 countries (some of them contributed to the previous CO_Sorghum version)
- 27 variables identified
  - 14 already in CO_324
  - 13 additional variables

=> It is only when people « get involved / are concerned / use the information » that it is efficient

Developping shared vocabularies
Requirements for interaction and decision tools

- A tool (website) to submit new variables to the crop ontology

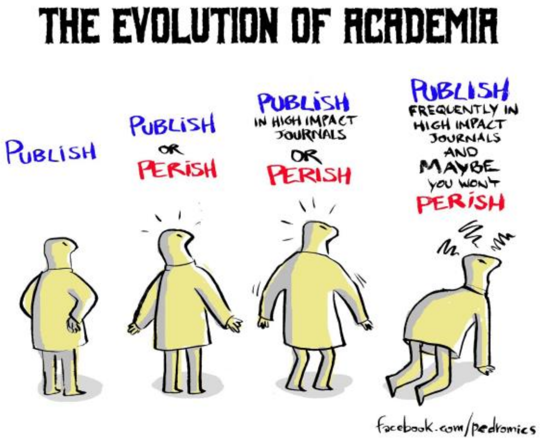
- A Git like tool to « exchange » on the variables / methods / scales
  - Relevance of adding new traits / variables / methods / scales ?
  - Obtain information on already available traits

- A group of plant experts / curators (involved in the concerned species) validating / challenging the new inputs ?

Developping shared vocabularies

Editor's support

- Science = papers….


THE EVOLUTION OF ACADEMIA
PUBLISH
PUBLISH OR PERISH
PUBLISH IN HIGH IMPACT JOURNALS OR PERISH
PUBLISH FREQUENTLY IN HIGH IMPACT JOURNALS AND MAYBE YOU WON'T PERISH
facebook.com/pedromics

- Science quality can be improved through the use of ontologies

- Providing ontologies for the variables in articles allows their re-uses, aggregations… Speed up research and allows reproductibility analyses

- Contacting editors to make relevant variable information compulsory !!

Developping shared vocabularies

A collaborative effort !

cirad
LA RECHERCHE AGRONOMIQUE
POUR LE DÉVELOPPEMENT

agap

AIDA
Unité de recherche
Agroécologie et intensification durable des cultures annuelles

Jean-François Rami

Sandrine Auzoux

Lauriane Rouan

iavao
Innovation et amélioration variétale en Afrique de l'Ouest

Elizabeth Arnaud

Marie-Angélique Laporte

Bioversity International

Crop Ontology
for agricultural data