



Process-sensitive naming: The matter of classification

Sabina Leonelli

Exeter Centre for the Study of Life Sciences

& Wissenschaftskolleg zu Berlin

@sabinaleonelli



Outline

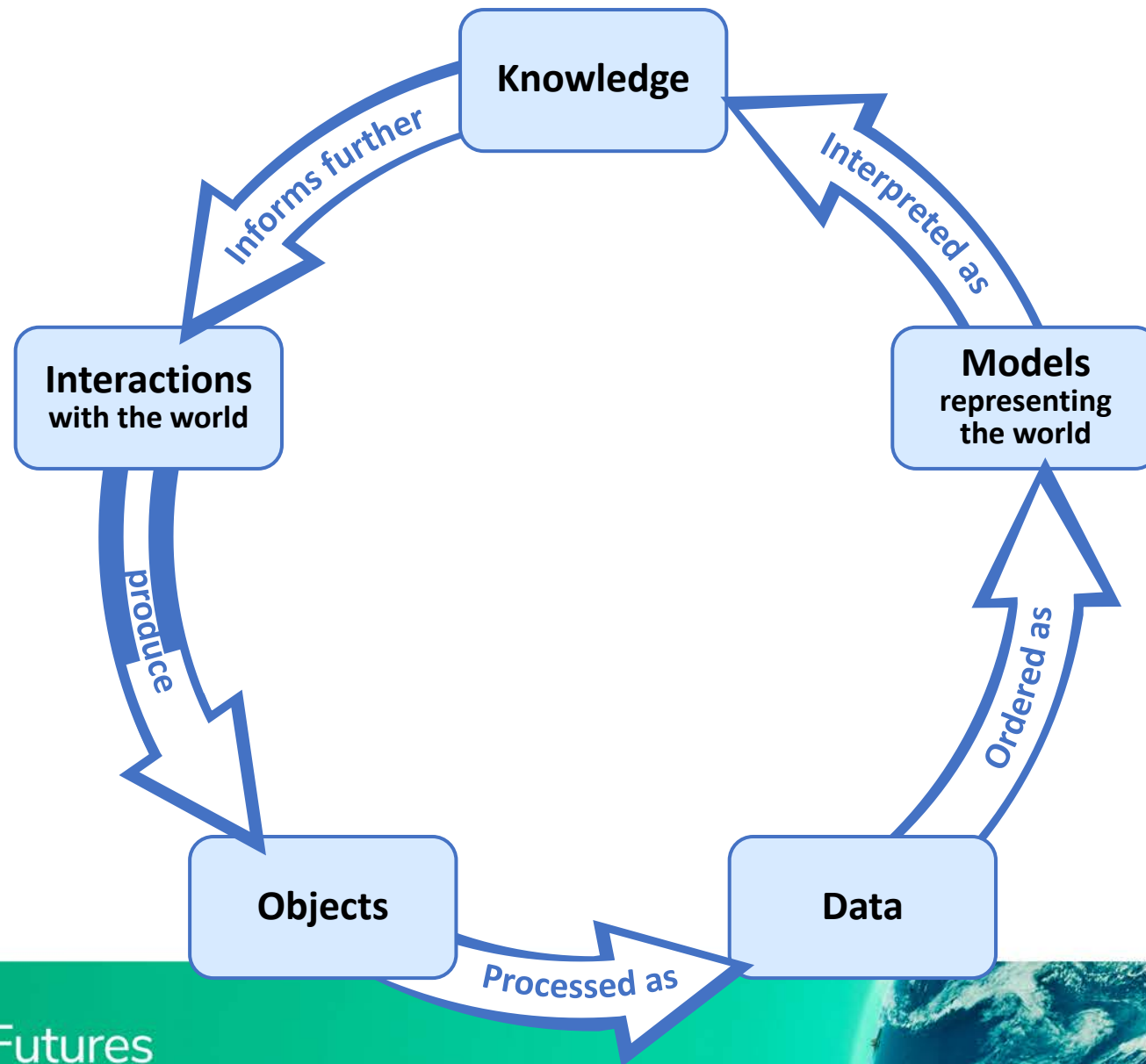
PART ONE: Classification in data systems (data semantics)

PART TWO: Data semantics in action: epistemic and social justice in crop data classifications

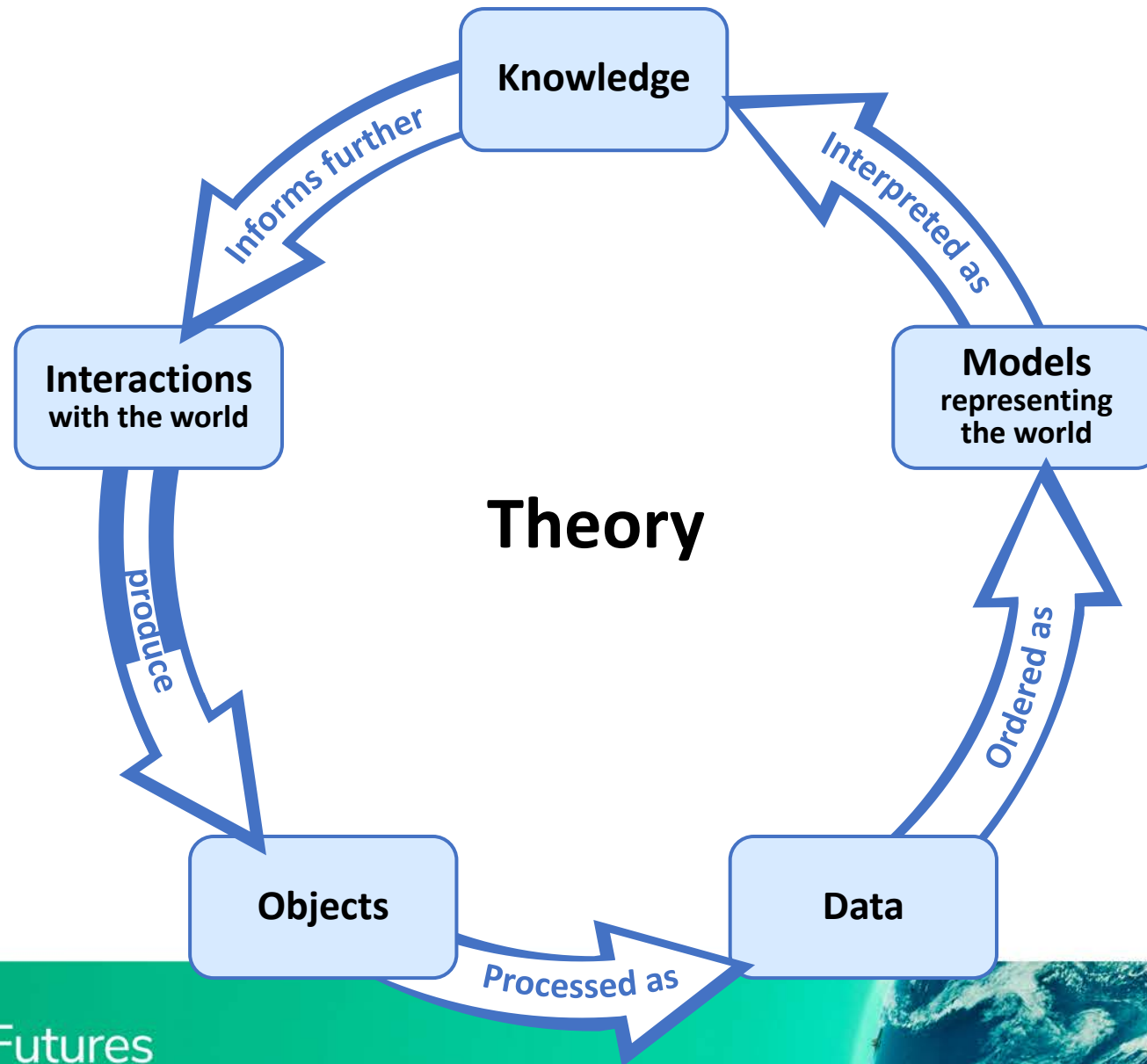


PART ONE – DATA SEMANTICS

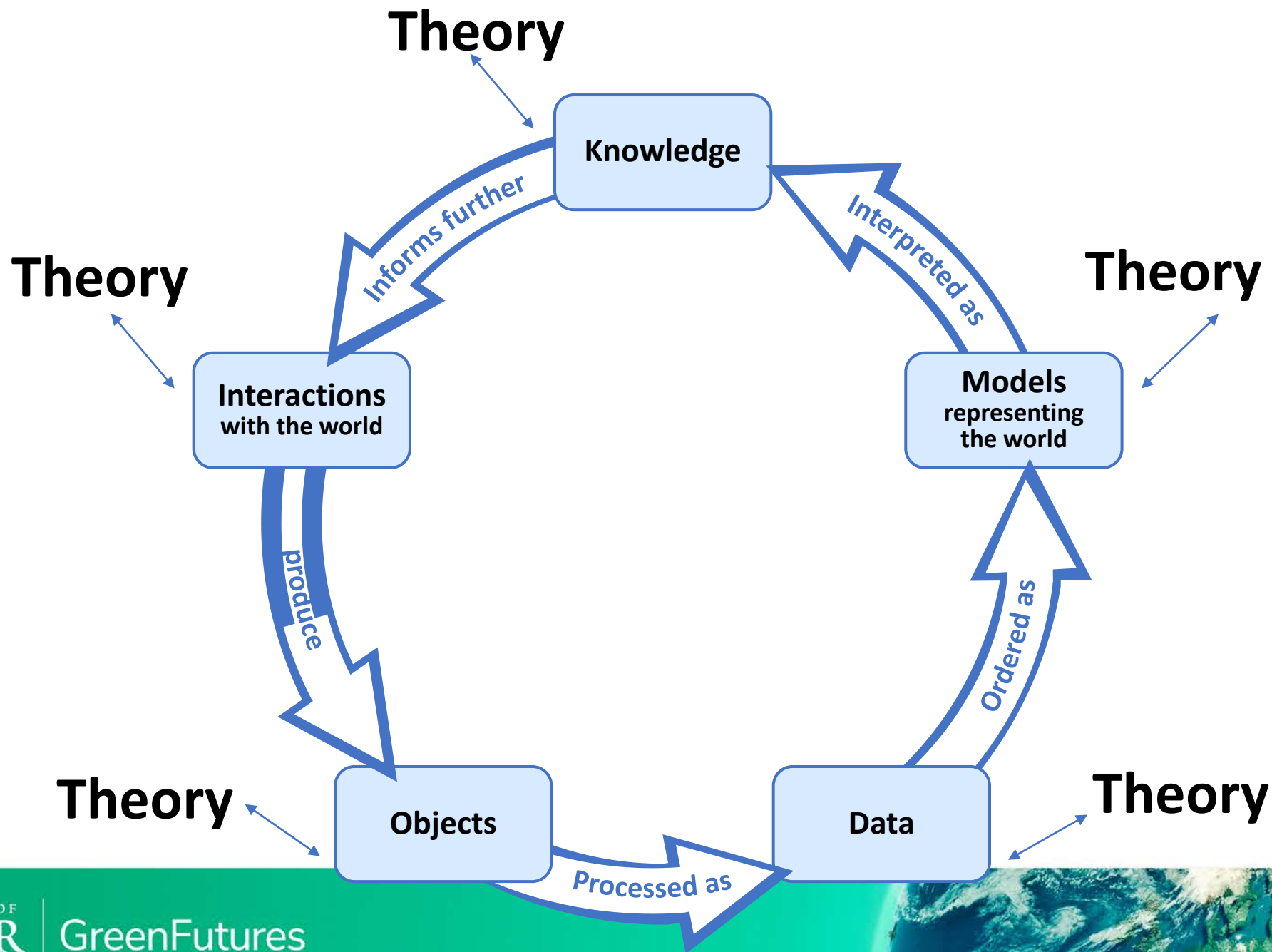




Leonelli 2018,
2019



Leonelli 2018, 2019



Data semantics: Oldest challenge, newest frontier

- Linkage of highly varied, culturally embedded terms associating data to phenotypic and genotypic traits, and biological phenomena of interest
 - For personalized medicine (e.g. clinical use of genomic data for cancer diagnostics)
 - For precision agriculture (e.g. integration of field trial results around the world with genomic data to help breeding choices and directions)



Crucial role of semantic interoperability

Data semantics systems determine

- how data are incorporated into machine learning algorithms;
- which data are linked with each other, and how
- which claims – and about what – data are taken as evidence for;
- whose knowledge is legitimised or excluded by data science & AI;
- whose perspective is incorporated in data-driven knowledge systems.

They inform key aspects of data science & its applications:

1. the choice of expertise and domains regarded as relevant to shaping data mining procedures and their results;
2. the development and specifications of data infrastructures, including what is viewed as essential knowledge base for data mining;
3. the governance of data dissemination and re-use through such infrastructures.



Classification, Unification and the Risks of Monism

- Standardisation, unification and monism: Trouble in principle..
 - developing a unique semantic system has long been an aspiration for archivists and librarians
 - standards are crucial motor of data interoperability (e.g. Google!)
 - however, short step from standardisation to classificatory monism
 - agreement on standards unavoidably involves loss of system-specific information that often matters to data interpretation



Classification, Unification and the Risks of Monism

- ..and trouble in practice
 - the variety of stakeholders, data sources and locations at play inevitably results in a proliferation of classification systems and increasing tensions among different interest groups around what system to adopt
 - standards help to make data classification user-friendly..
 - ..*for some!* Huge issues with exclusions, demarcation of expertise and negotiating boundaries around research work
 - Also, evidence that data re-use often linked to participation in *developing* data infrastructures



Classifying to integrate: how?!

- The Linnaeus model: centralization with an emphasis on progressive expansion and revision
 - E.g. taxonomy; medical thesaura; model organism research
- The EU model: decentralization and federation through interoperability standards and guidelines
 - Examples: the European Open Science Cloud



Towards *problem-driven and value-laden data semantics*

Crucial to develop interoperable data semantics that takes account of

- **Specificity**
- **Diversity**
- **Degrees of entrenchment &**
- **Inclusivity**

of relevant systems of practice



PART TWO – Fostering epistemic and social justice in crop data classifications



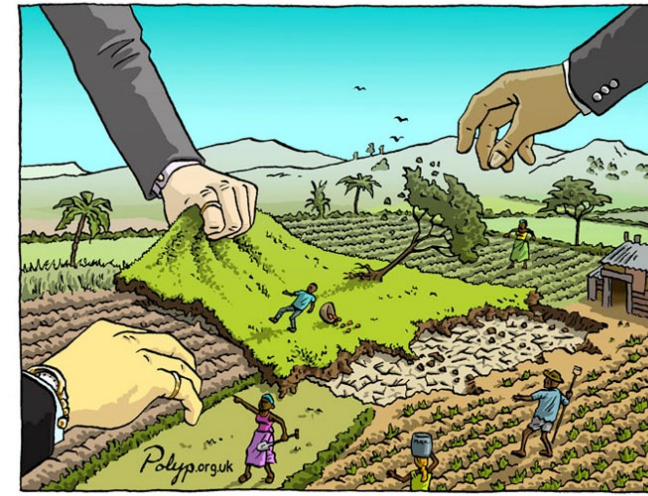
Crop data sharing dreams..

- Global plant knowledge as *common good* for planetary health
 - re-imagining agriculture away from high-yield monocultures
 - understanding bio/agrodiversity to boost sustainable use of plants
- Strong push to collect and share
 - Phenotypic data, including roots (e.g. use of radar for time series phenotyping) and remote sensing (satellite, drones)
 - Breeders insights and observations
 - Traditional/local knowledge
 - Agricultural technologies as data collecting vehicles
- Challenges to understandings of ethnobotany as extraneous to scientific research
 - aim to incorporate insights – and, crucially, datasets – acquired through local knowledges into an internationally-recognized evidence base for agronomy and plant science



.. and nightmares

- Benefit-sharing problems:
 - Market control and IP regimes by aggrotech: data sharing sits alongside centuries-long exploitation and commodification of farming
 - New recognition of phenotypic data does not extend to the communities who create them
 - Debate around who should reap the benefits of such research – and particularly whether traditional communities should partake in its development and profits - continues with no resolution
- Surveillance concerns:
 - unregulated mass surveillance of farming behaviors and local environmental characteristics



.. and nightmares

- Molecular bias endures:
 - Molecular data consistently rated as more reliable than observational accounts of the physiology and development of morphological traits, and therefore used to structure databases and data analysis
 - Researchers/contributors with no related expertise and instruments are viewed as second class citizens
- Digital divide expands:
 - Access to latest digital technologies and related infrastructures viewed as essential to research



Monism in crop data sharing

- Market and national agricultural policies overdetermine data economy and scientific/evidential use of data
- Data sharing and standardisation: loss of environmental / material / cultural context
- Flattening of epistemic space: regime of agricultural development erases cultural, biological, scientific and semantic diversity

[Scott, Harwood, Bonneuil, Fullilove, Saraiva, Curry.. Leonelli 2022a/b]



Semantic systems and the role of community science: the case of cassava





**RESEARCH
PROGRAM ON
Roots, Tubers
and Bananas**



Trait Descriptors

- Fixed hierarchical organization
- Focus on
 - Taxonomic identification before entering collections as an accession
 - Characters associated with “reference specimen” rather than capturing environmental variation
- Limited use to breeders interested in multi-site evaluations under different management practices and environmental stressors

[FAO/CGIAR 1970-2000, Leonelli 2022, Curry & Leonelli under review]

Table 2. Descriptors categories and their classes used in the morphological characterization of cassava germplasm, Chapadinha, MA, 2013.

S/N	Plant descriptors	Given categories
1	Branching habit	1-Erect; 2-Dichotomous; 3-Trichotomous and 4-Tetrachotomous.
2	Type of plant	1-Open; 2-Umbrella type and 3-Compact
Leaf descriptors		
3	Apical leaf color	1- Light green; 2- Dark green; 3- Purplish-green and 4- Purple.
4	Pubescence of apical bud	1-Present and 2- Absente.
5	Petiole color	1-Yellowish-green; 2- Green; 3-Redish-green; 4- greenish-red; 5-red and 6-Purple
6	Developed leaf color	1-Light-green; 2-Dark-green; 3-Purplesh-green and 4-purple
7	Terminal branches color	1-Light-green; 2-Dark-green; 3-Purplesh-green and 4-purple
8	Leave’s rib color	1- Green; 2-Redish-green; and 3- Greenish-red
9	petiole position	1-Tilted up; 2-Horizontal; 3-Angled down and 4-Irregular
10	Prominence of leaf scars	1-Without prominence and 2-Proeminent.
Stem Descriptors		
11	Color of stem cortex	1-Light yellow; 2-Light green; 3-Green and 4-Dark green.
12	Length of phyllotaxis	1-Short; 2-Middle and 3-Large.
13	External Color of steam	1-Orange; 2-Yellowish-green; 3-Golden; 4-Light brown; 5-Gray; 6- Silvery; 7- Gray; 8- Silvery; 9- Dark brown.
14	Color of stem epidermis	1- Cream; 2- Light brown; 3- Dark brown; 4- Yellow.
15	Growth habit of the stem	1-Straight and 2-Forked.
Root descriptors		
16	Presence of peduncle in roots	1-Present and 2-Absent.
17	External color of roots	1-White; 2-Yellow; 3-Light brown; 4-Brown and 5-Dark brown.
18	Color of root Cortex	1-White; 2-Yellow and 3-Pinkish.
19	Texture of root epidermis	1-Smooth and 2-Rough.
20	Constriction of roots	1-Absent; 2-Little or none and 3-Average.
21	Root shape	1-Conical; 2-Cylinder and 3-Spindle.
22	Highlight pellicle from roots	1-Easy release and 2- Difficult release.
23	Highlight of roots cortex	1-Easy release and 2-Difficult release.
24	Position of roots	1-Horizontal and 2- Vertical tendency.

Challenges from the field/I

- **Biodiversity:** Variability within and across crop types
 - What counts as an individual, a variety, an accession, an ecotype, a strain (especially for clonal reproduction)?
 - Soil, environment, nurture all strongly affect phenotype: how to pick out representative/‘normal’ traits?
 - Which individuals are representative of a plot?



Challenges from the field/2

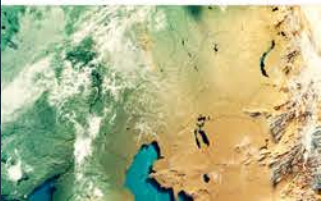
- **Methodological diversity:**

Variability in skills and measuring methods for traits data collection

- how plot is measured and sampled; color assessment; how harvest is collected (the 'standard cut'); counting flowers

Shifting temporalities of data collection

- Time of flowering and harvest can vary considerably within and across trials
- Constant monitoring and no standardization
- Processing and management of germplasm materials



Challenges from the field/3

• Cultural diversity:

Naming of varieties

- Locally named from person who gives the clone, so breeders need to reconstruct clone origin by tracking back the gifting of clones
- Genetic markers used to determine 'real' name – however, unclear taxonomy of phenotype, and big implications for local culture/markets
- Uneasy relationship to traditional taxonomy

Identification of valuable traits: which trait matters and for whom?

- Trialing varieties can take 6-7 years, crucial to consult with farmers regularly
- Gender differences & complex local mediations around choice of variety for any given trial



Challenge of instituting global standards for local, situated procedures



Problems with IPGRI descriptors: structural stability and narrow focus

- What makes them effective standards and benchmarks for researchers of different backgrounds looking to identify a given variety and validate its taxonomy before entering it into in situ or in vitro collections
- But prevent these descriptors from being able to capture:
 - the biological diversity exhibited in the countless, variously adapted and constantly evolving forms of plant life,
 - the scientific diversity in the methods and skills used by data collectors responsible for measuring and implementing descriptors in the field,
 - the cultural diversity manifested in existing ideas around what constitutes a valuable trait.

IPGRI descriptors are therefore of limited use to researchers studying plant environmental responses and breeders aiming to test crop varieties in multi-site evaluations and under different environmental conditions and management practices.



Crop Ontology on cassava: 365 traits in 2021 (from 120 in 2011)

Ontology browser

Search and browse ontologies

Find exact ID Find |

Search for text CO (Cassava Trait Ontology)

- CO:0000000 CGIAR cassava trait ontology
 - + is_a CO:0000001 Agronomic trait
 - + is_a CO:0000002 Morphological trait
 - + is_a CO:0000003 Physiological trait
 - + is_a CO:0000004 Quality trait
 - + is_a CO:0000005 Stress trait
 - + is_a CO:0000006 abiotic stress
 - + is_a CO:0000007 biotic stress t
 - + is_a CO:0000027 bacterial d
 - + is_a CO:0000029 fungal dise
 - + is_a CO:0000038 Cassa
 - + is_a CO:0000032 Cassa
 - + is_a CO:0000030 insect dan
 - + is_a CO:0000028 viral disea
- + GO:0008150 biological_process
- + GO:0003674 molecular_function
- + GO:0005575 cellular_component
- + PO:0025131 plant anatomical entity
- + PO:0009012 plant growth and developn
- + SO:0000400 sequence_attribute
- + SO:0001060 sequence_variant
- + SO:0000110 sequence_feature

CO:0000032 'Cassava anthracnose disease severity'

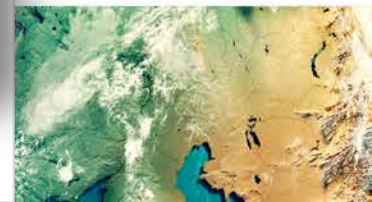

Cvterm details

Term id **0000032**
Term name **Cassava anthracnose disease severity**
Term definition **Severity of the Cassava anthracnose disease caused by Colletotrichum gloeosporioides f. sp.**
Comment **Symptoms: Cankers on the stems and bases of leaf petioles.**

Synonyms
"Cassava anthracnose disease severity - Cassavabase" EXACT []
"CADSev" EXACT []
CADS EXACT
"CADS" EXACT []

Definition dbxrefs
CO:curators

Definition dbxrefs
TO:0000439



Courtesy of Afolo Agbona, Peter Kulakow and Elisabeth Arnaud, IITA/CGIAR, Crop Ontology

Stem Color
CO_334:0000062

Root Shape
CO_334:0000020

Root Neck Length
CO_334:0000022

Outer Skin Color
CO_334:0000064

Root Number
CO_334:0000011



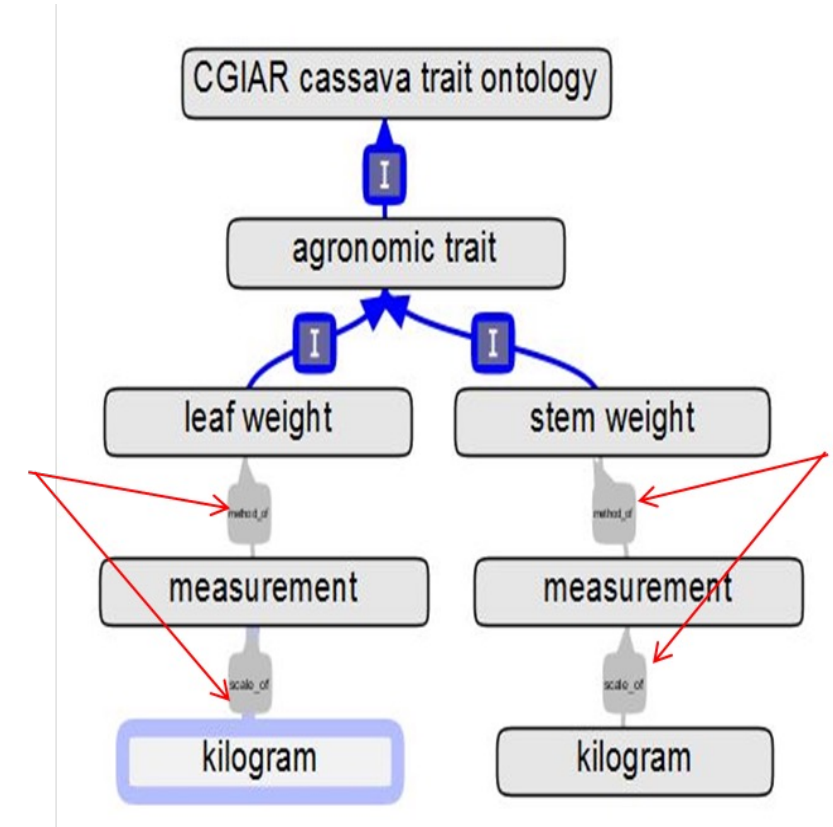
Process-sensitive naming

- **Capture human interactions with plants and their environment, rather than plants by themselves**
- Naming processes not solely as application of nomenclature, but as explicitly tied to methods of data collection *and* processing
- Facilitates finer-grained look at *intra-species* diversity
- Metadata focus on documenting difference as well as similarity and problematizing typicality of traits vis-à-vis species concepts
- Spurs efforts to document history of data processing
- **Environment takes center stage: from focus on trait in itself to *relation between trait and context***

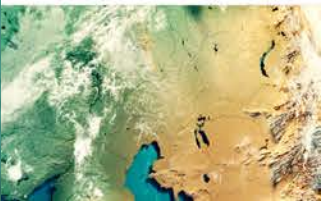


Process-sensitive naming: capturing *methodological* variation

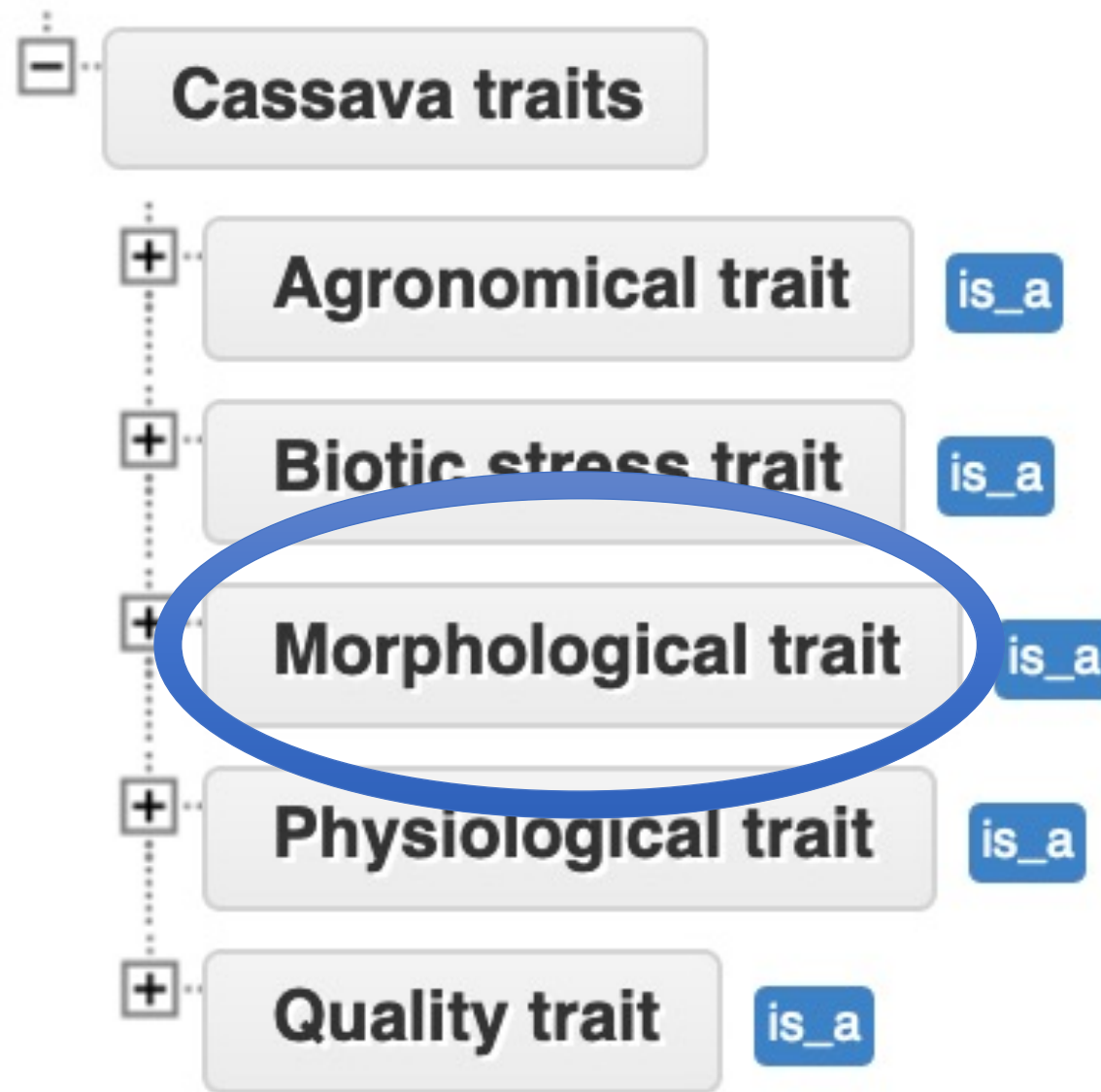
- Main relation among traits is ‘**variable of**’
- Traits classified via Definition, Measurement Method and Scale (‘scale_of’, ‘measured_by’)
- **Can deal with variation in measurement approaches and skills**
 - multiple shades of colors: captured through changing skill (so color is still yellow overall, but one can use skill of discerning color)

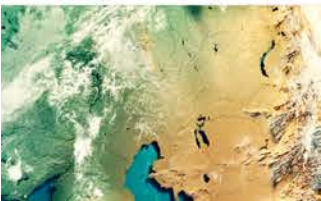
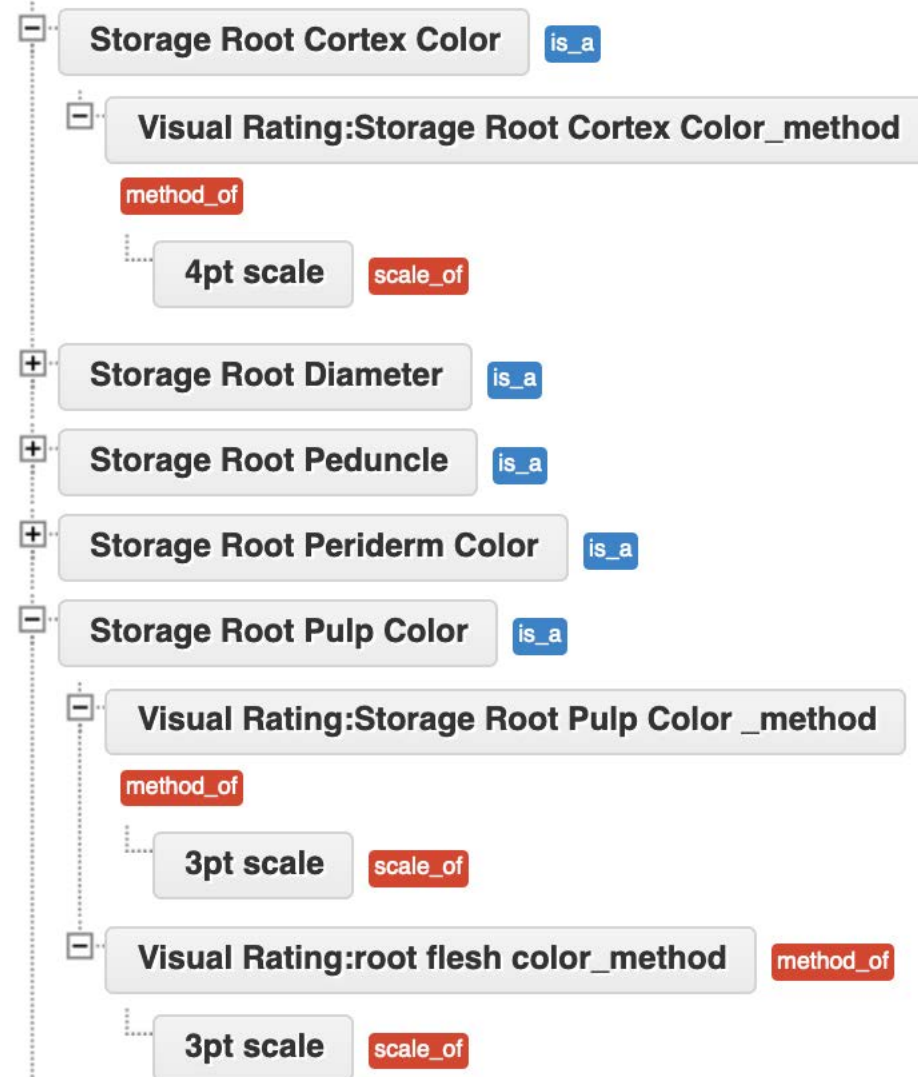
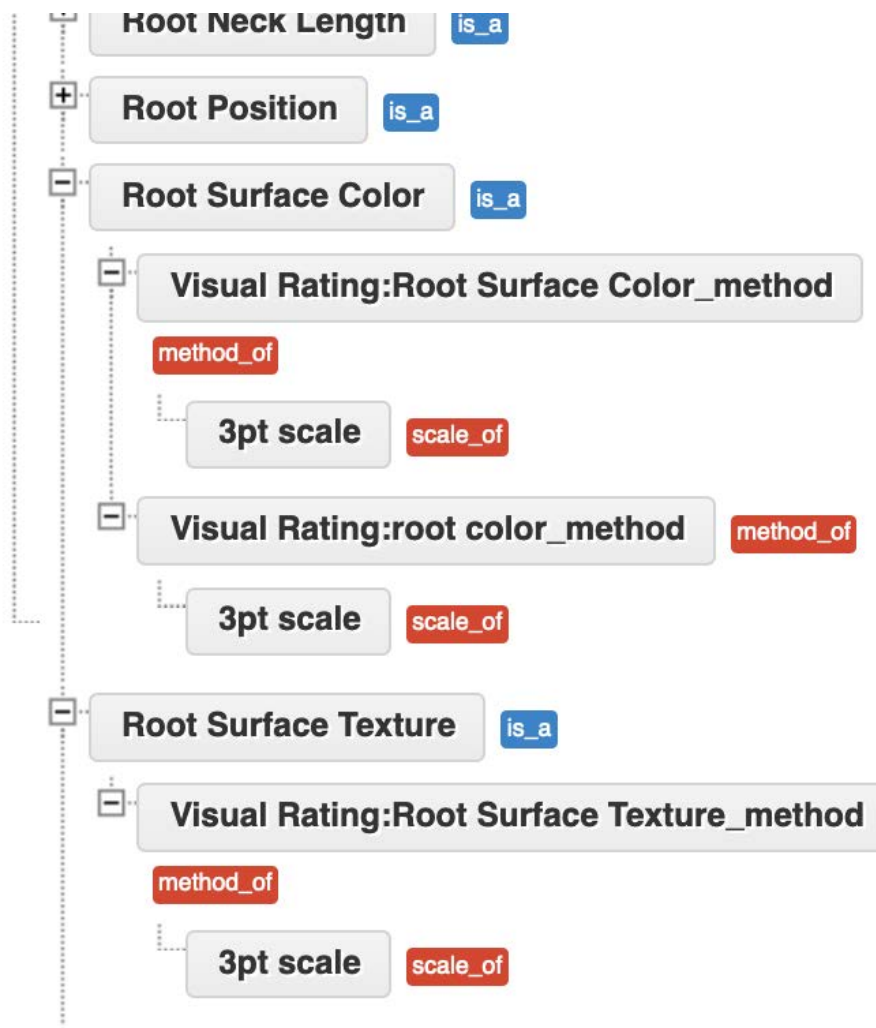


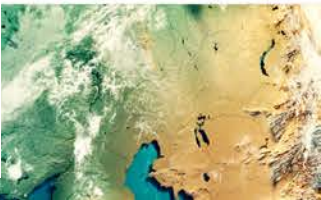
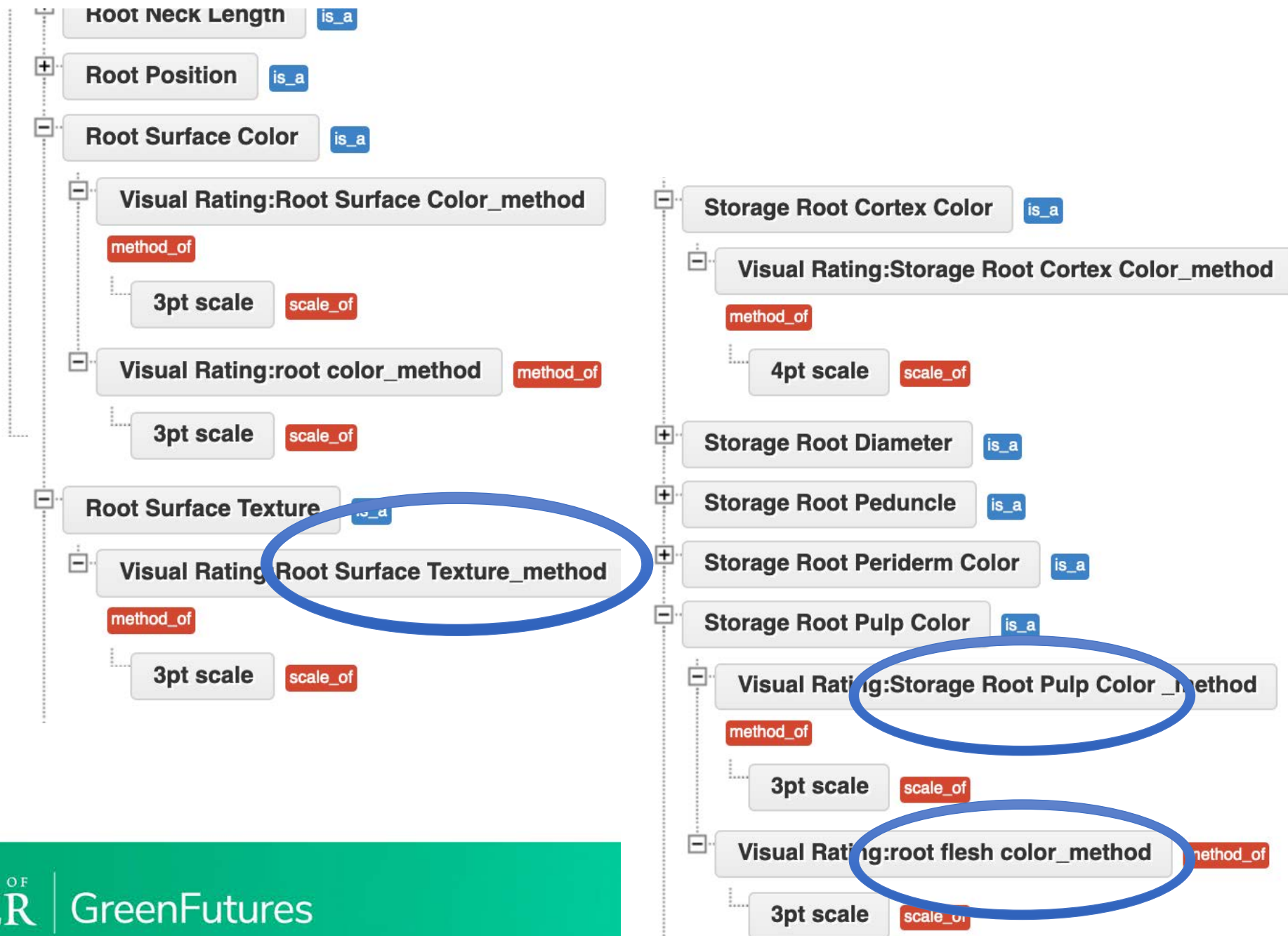
Leonelli forthcoming











Navigation

Term, Trait, Method, and Scale

- Ploidy
- Presence of pollen
- Proportion of female flower
- Proportion of plant with inflorescence
- Root Neck Length
- Root Position
- Root Surface Color
 - Visual Rating:Root Surface Color_method
 - 3pt scale
 - Visual Rating:root color_method
 - 3pt scale
- Root Surface Texture
- Root constrictions number
- Rotted Storage Roots
- Seed Color
- Sepal Color
- Stem Color

Variables

Term details

Key	Value
method_id	CO_334:0010436
method_name	Visual Rating:root color_method
ontology_id	CO_334
ontology_name	Cassava
method_class	Estimation
method_description	Visual rating of storage root surface color as evaluated by CIAT
language	EN
created_at	2021-09-28-10:36:32

category_1	1 = white
category_2	2 = intermediate
category_3	3 = dark brown

Navigation

Term, Trait, Method, and Scale

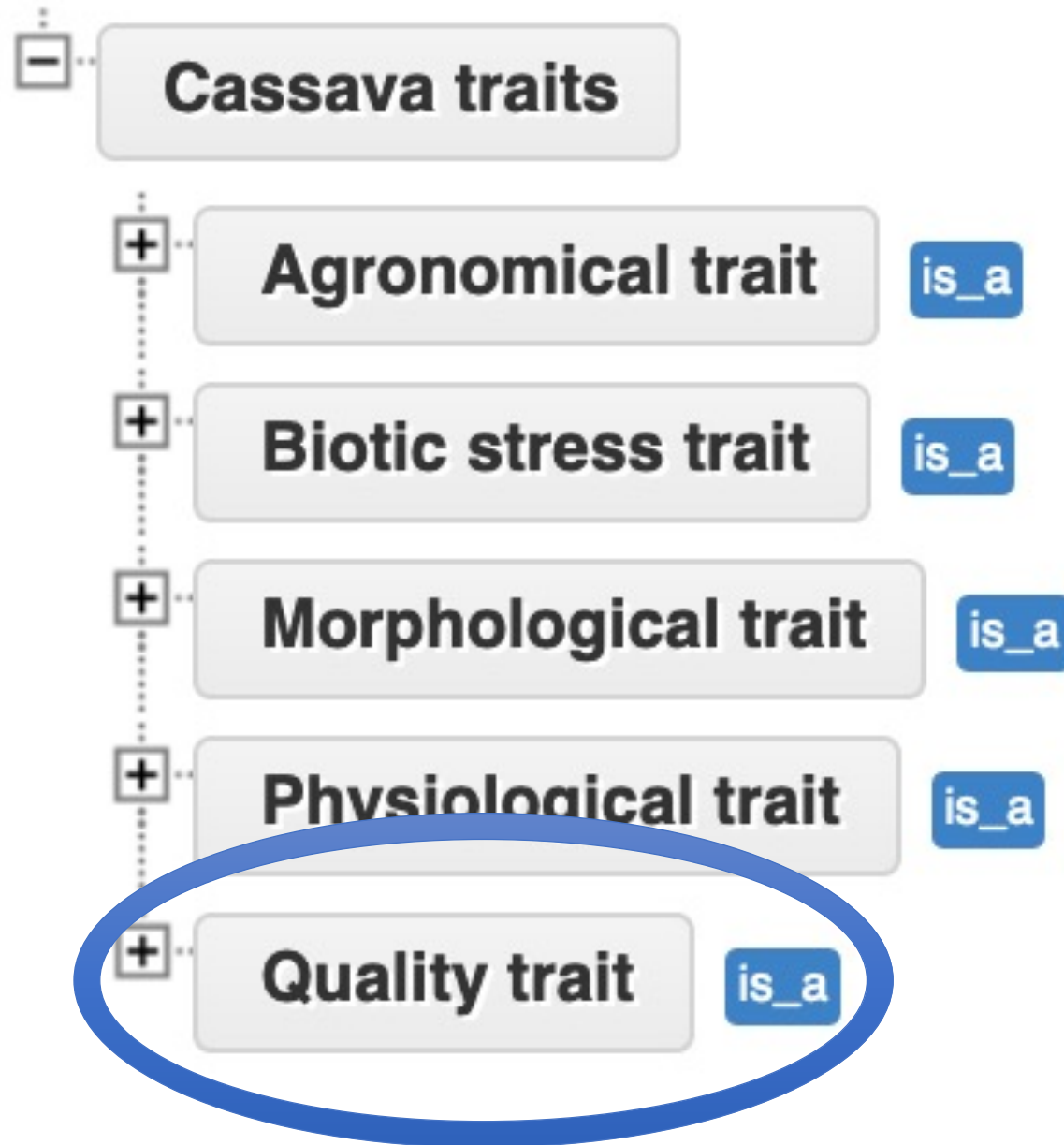
- Ploidy
- Presence of pollen
- Proportion of female flower
- Proportion of plant with inflorescence
- Root Neck Length
- Root Position
- Root Surface Color
 - Visual Rating:Root Surface Color_method
 - 3pt scale
 - Visual Rating:root color_method
 - 3pt scale
- Root Surface Texture
- Root constrictions number
- Rotted Storage Roots
- Seed Color
- Sepal Color
- Stem Color

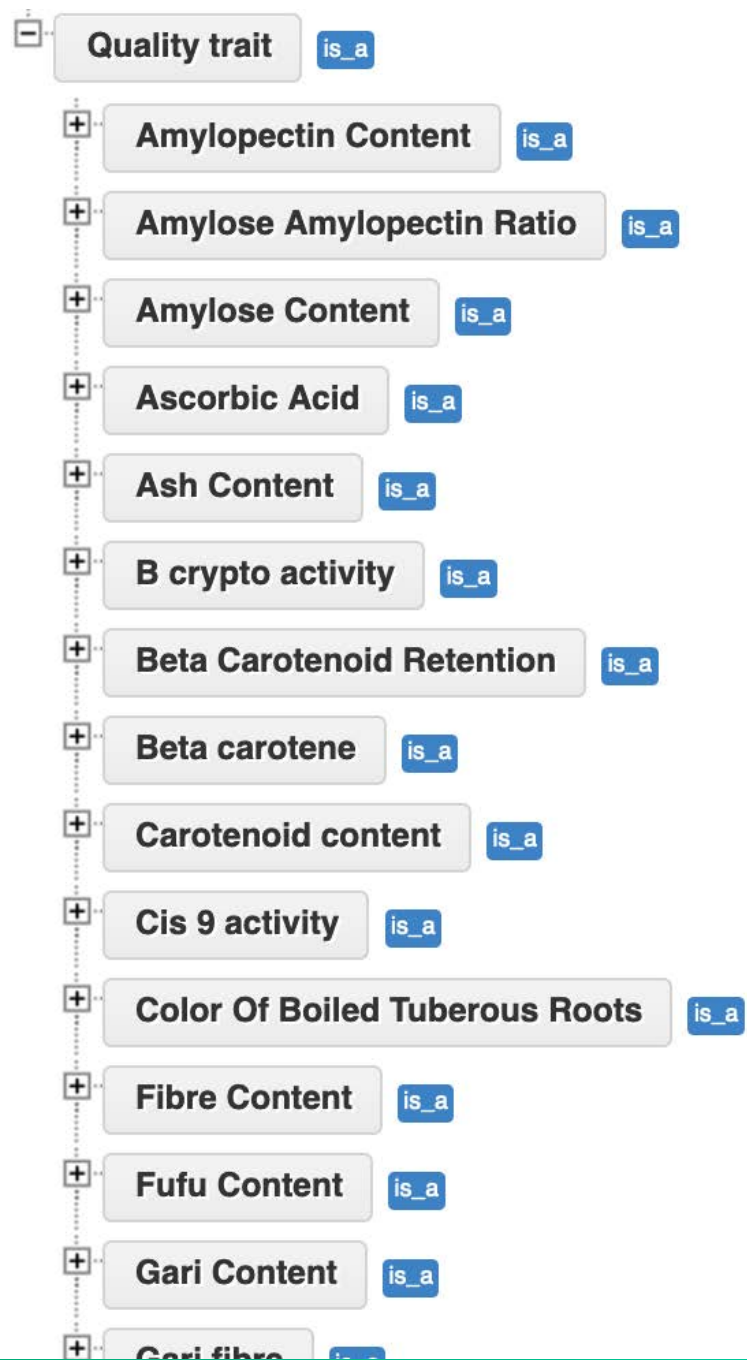
Variables

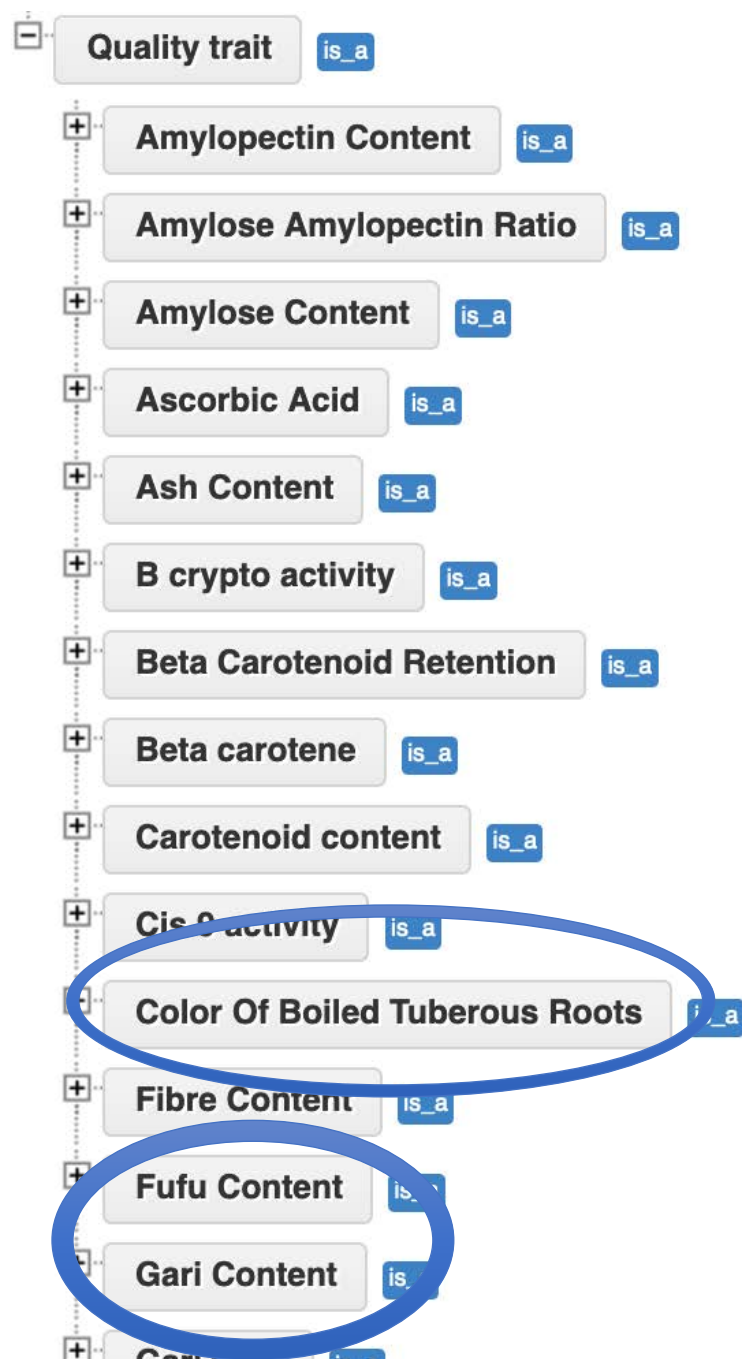
Term details

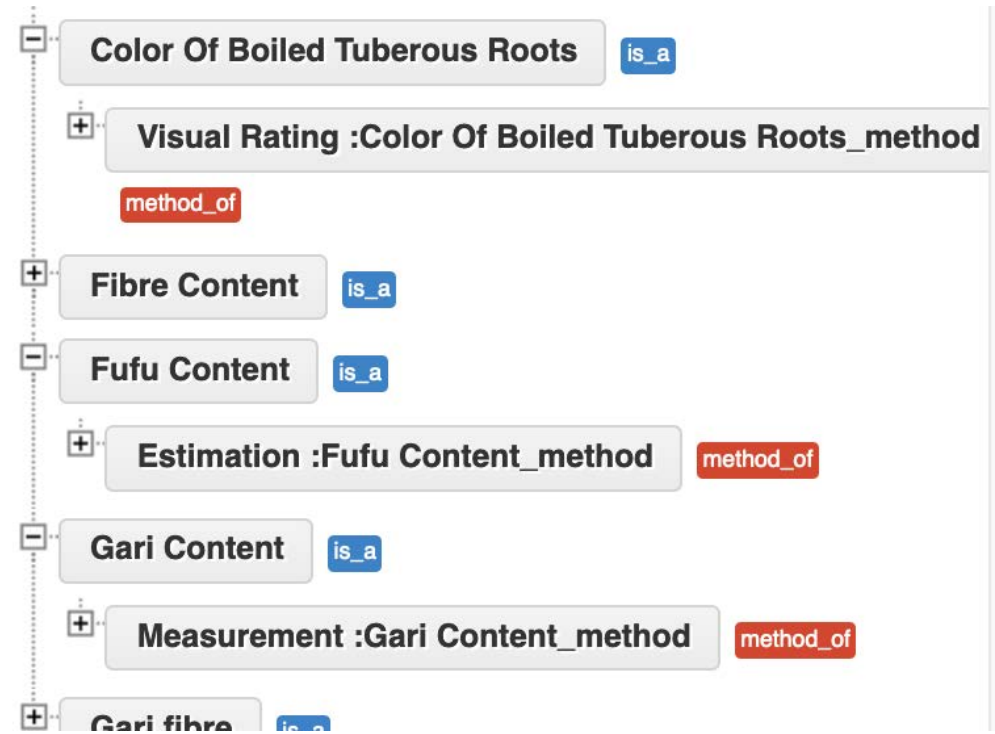
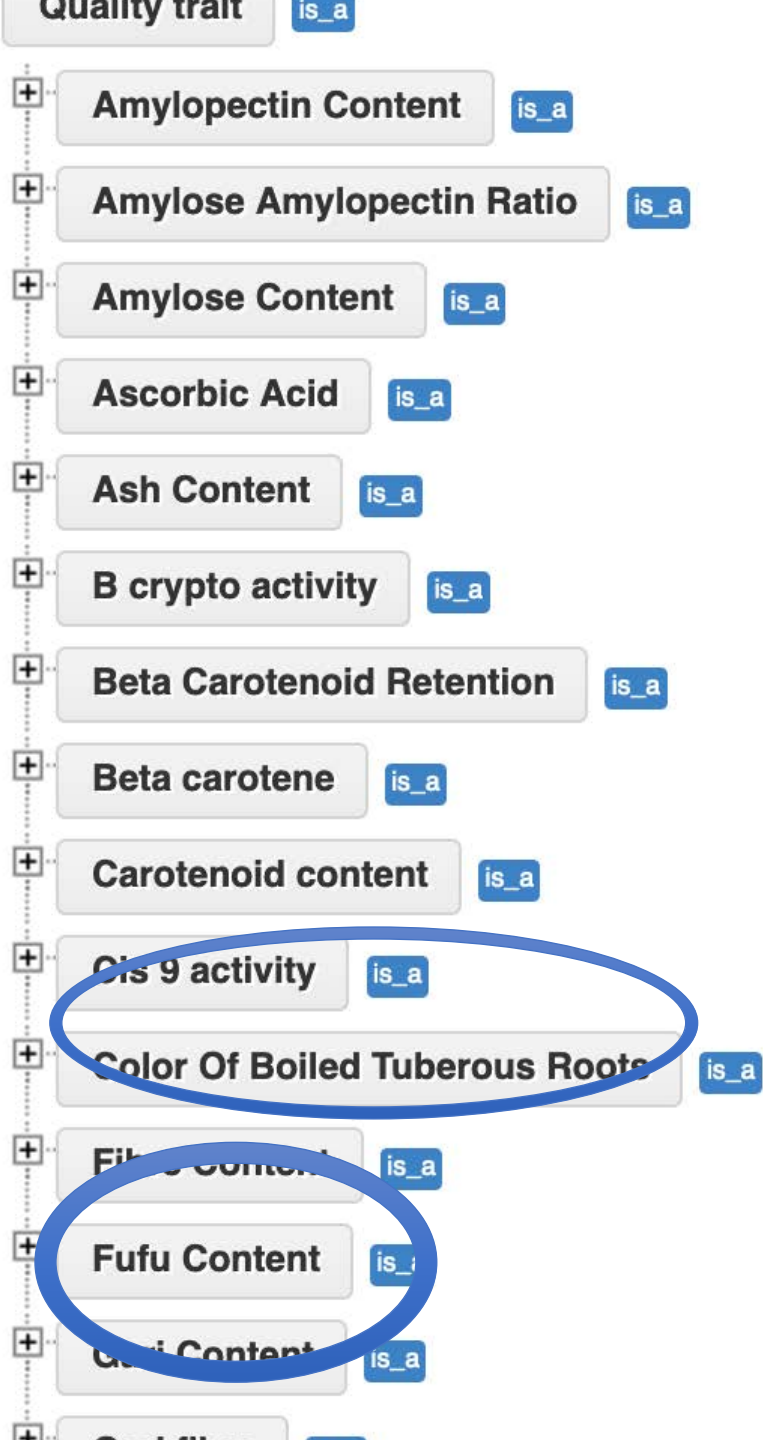
Key	Value
method_id	CO_334:0010311
method_name	Visual Rating:Root Surface Color_method
ontology_id	CO_334
ontology_name	Cassava
method_class	Estimation
method_description	Visual rating of root surface color
method_reference	Dixon et al 2010 Fukuda et al 2010
language	EN
created_at	2021-09-28-10:36:32

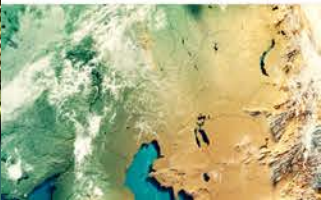
ontology_name	Cassava
scale_class	Nominal
category_1	1 = White or cream
category_2	2 = Light brown
category_3	3 = Dark brown

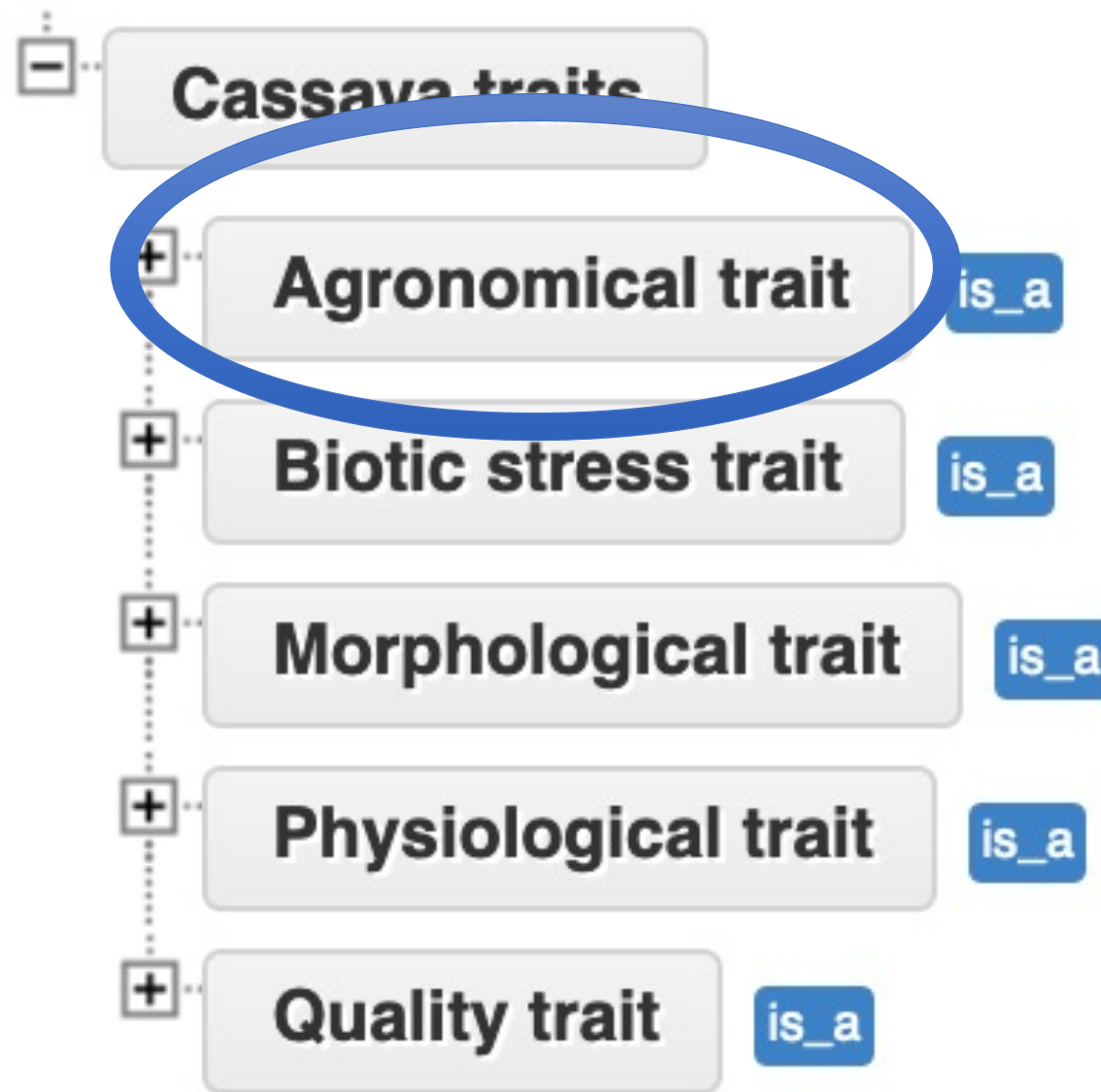


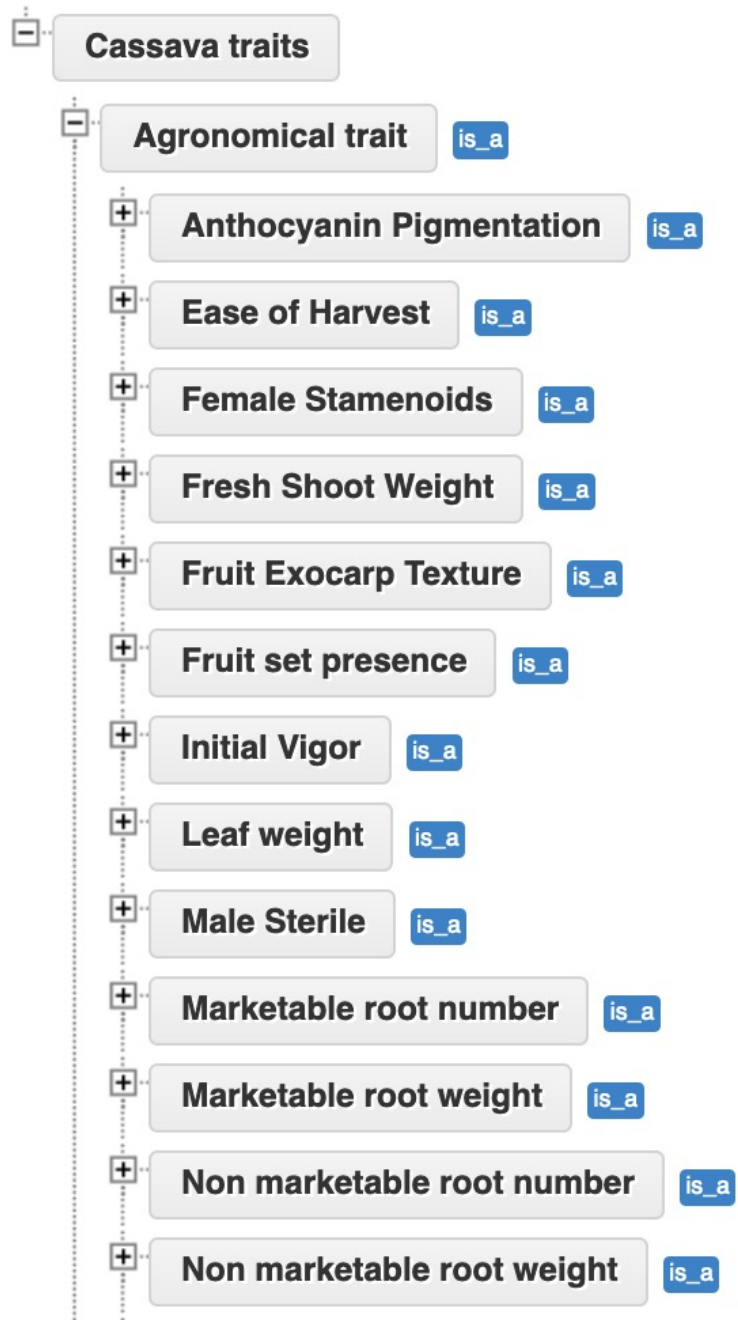


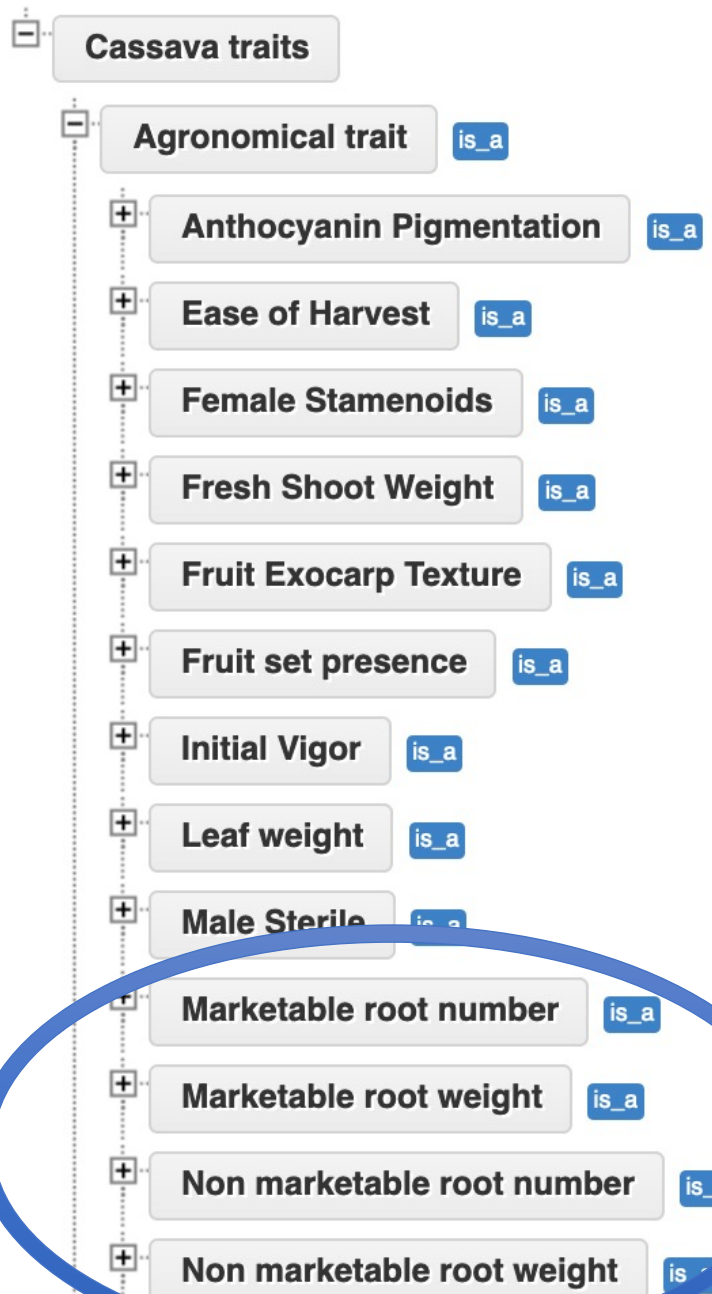


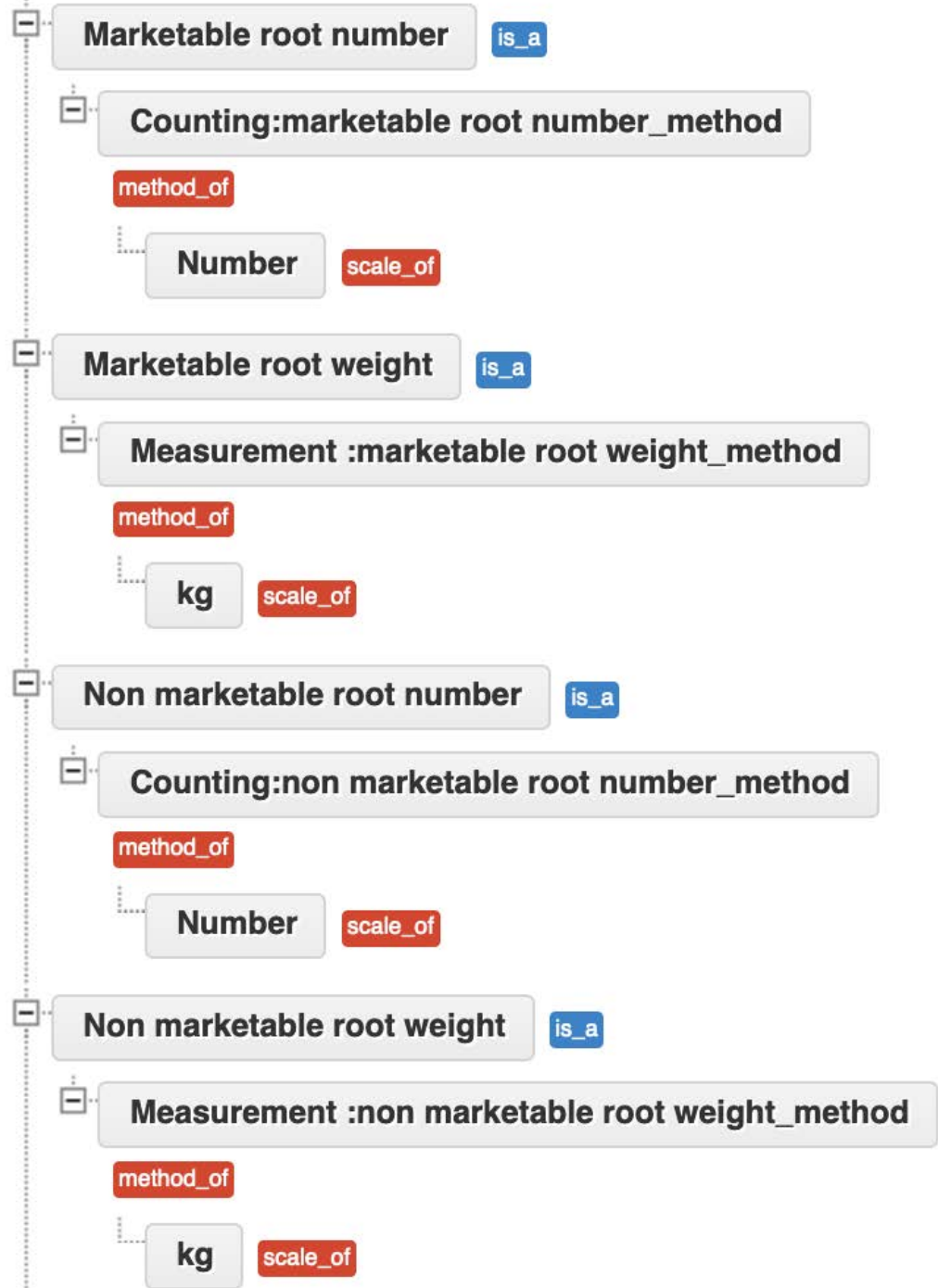












Process-sensitive naming: capturing *social* variation

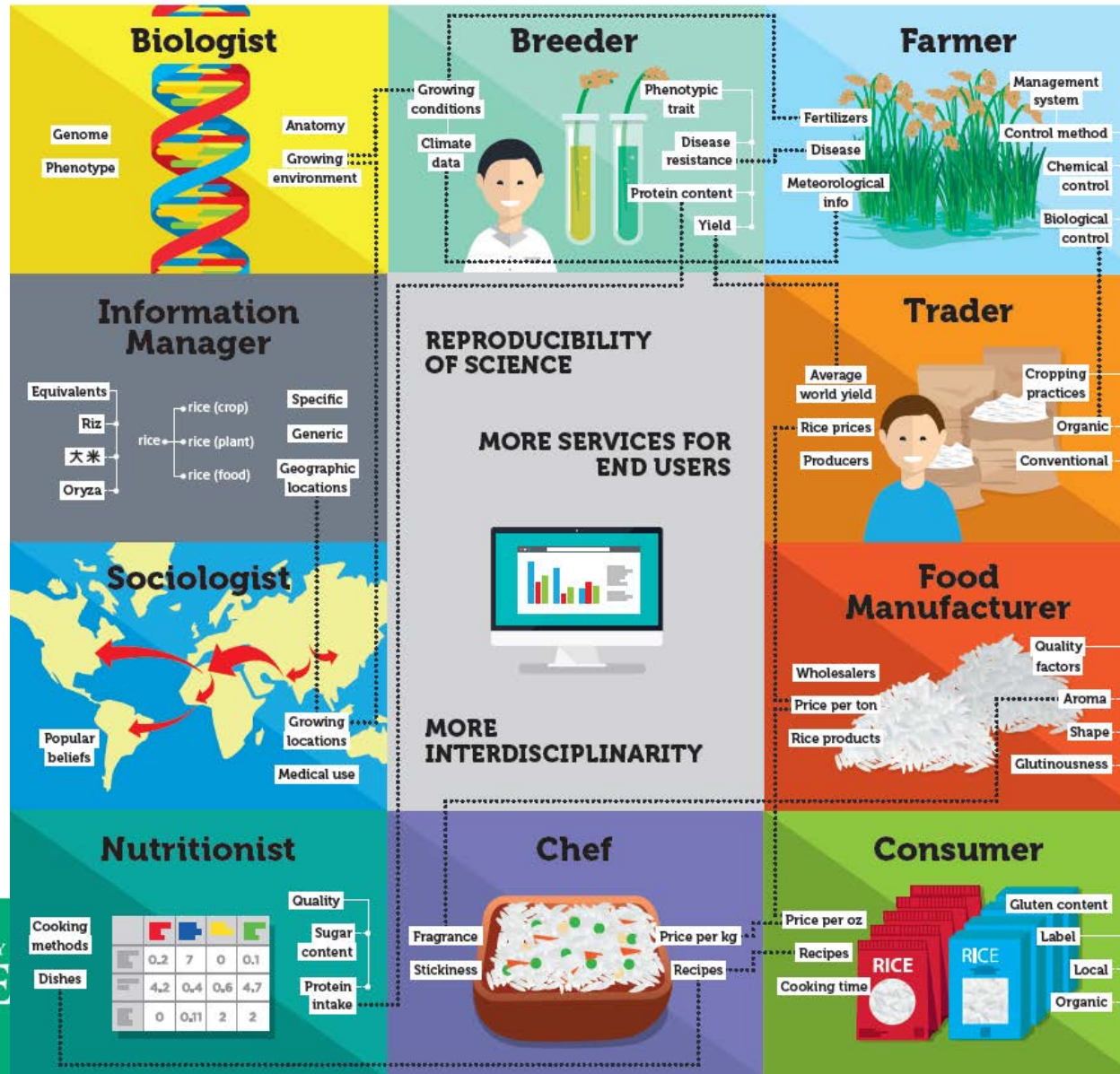
- “Ontologies need to capture everything people are doing, all the methods, there is no wrong or right way” (PI_17_A)
 - “If you get an accession, you should trace its history, get its attributes, in which trials it has been used and its performance in the trials at every level. Quality, agrobiotics, stresses. All information should be linked to accession identifier” (crop database curator)
- Critical role of community science: Assessing the need to add terms through open communication with breeders
 - Yearly farm visits and dialogue over preferred traits; Cassava breeders meetings and training sessions. E.g. what is “shoot weight”?
 - Use of English vs own-language descriptions → translation as significant in shaping variation of descriptions from user to user
 - Strong emphasis on gender-sensitive participatory evaluation
 - Direct farmer evaluation of varieties: triadic comparison of technology options (tricot)

[Leonelli forthcoming; Williamson and Leonelli (2022) *Towards Responsible Plant Data Linkage*. Springer Open Access.]



SEMANTICS - THE WAY TO RECONCILE POINTS OF VIEW AND DATA

THE EXAMPLE OF "RICE"

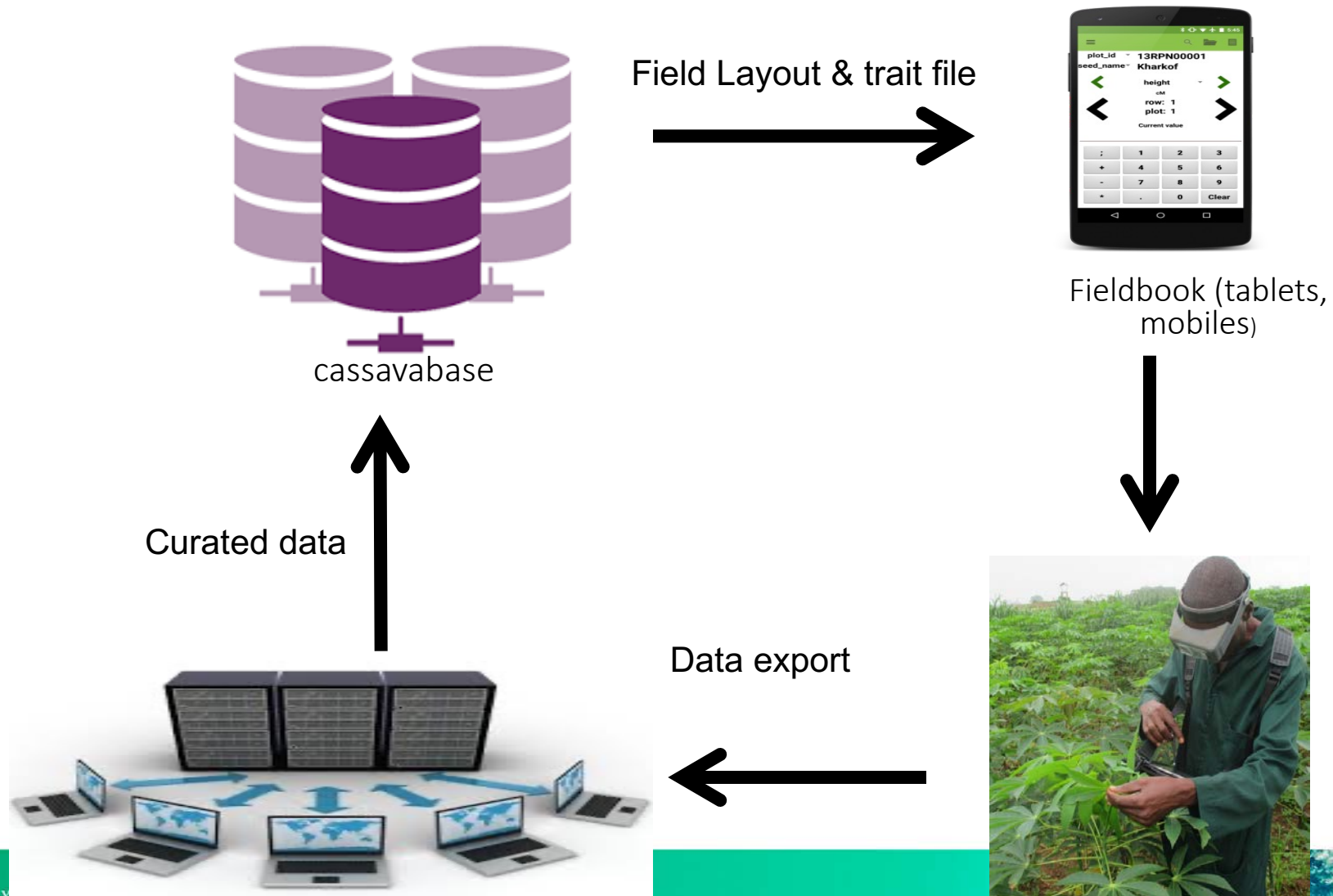


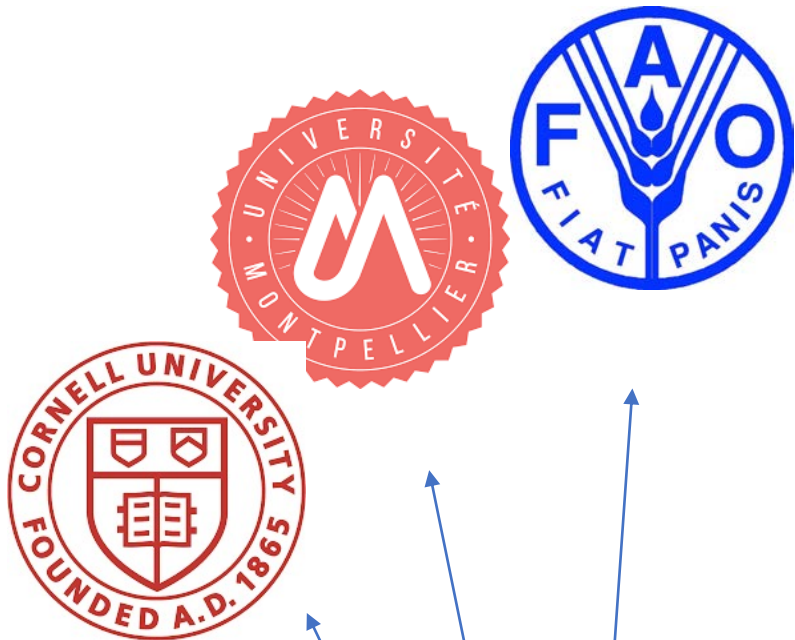
Research Data
Alliance(RDA)
Agrisemantics
Working
Group,
2017

Addressing Epistemic Injustice: Crop Ontology as Transnational Data Broker

More effective than trait descriptors in mediating national and international coordination & inclusion of multiple stakeholders:

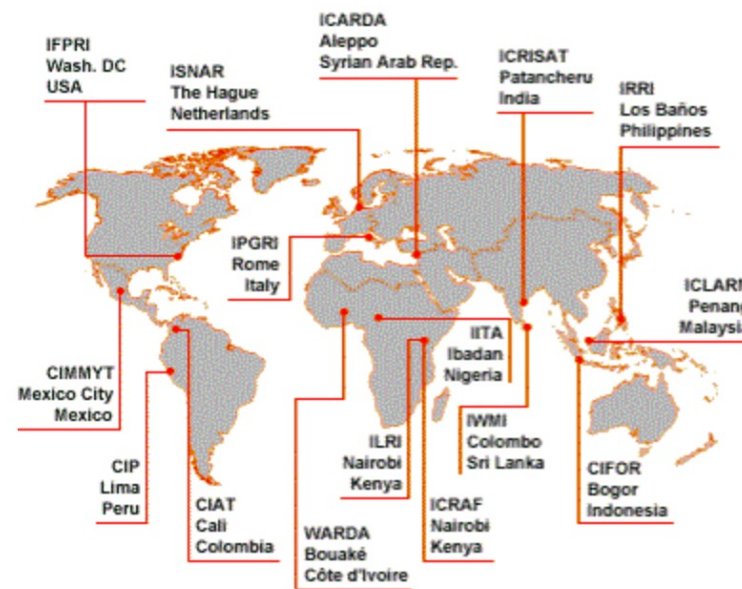
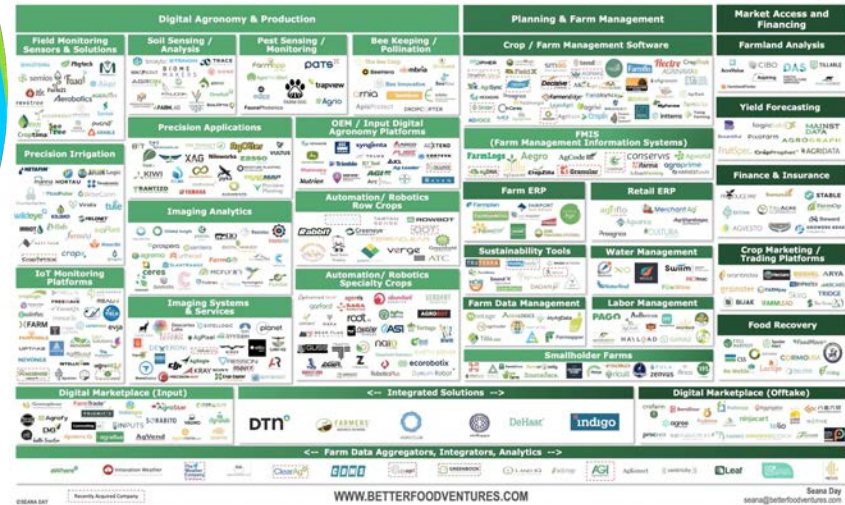
- **develop** participatory practices for crop data collection and management
 - involving local breeders and farmers into the development of data infrastructures, addressing issues such as gender inequality in their protocols and training, and resisting privatization of the resulting data through commitment to Open Data principles
- **mediate** between crop-specific, local databases and international initiatives in plant data management
- **negotiate** the tensions arising from attempts to link locally acquired digital information into global networks, and the related effort to regulate the transfer of plant genetic materials, such as germplasm, across national borders





BAYER

FARMTECH LANDSCAPE 2020



Addressing Social Injustice: Overdetermination by political economy?

Specific epistemic economy:

- Ecumenical attitude incorporating diverse value systems; disregard for borders (national or otherwise)
- More perspectives and expertise included in data systems, enabling more participation

But under overarching regimes of agricultural development, epistemic justice comes at a price..

- Enshrines data reuse as South to North
 - Little benefit to contributors
 - Unchallenged downstream commodification
- Laden with commitment to high-yield agriculture
 - Focus on new hybrids and “precision monocultures” (‘expert’ approach to food security)
 - Dismissive of subsistence agriculture and other models

[Leonelli 202, forthcoming; Curry and Leonelli forthcoming; Williamson and Leonelli forthcoming, under review]



Conclusion: Enduring tensions between epistemic and social justice in data classification

- Data classification is a crucial site for epistemic negotiation and epistemic justice
- Importance of community science efforts..
 - Process-sensitive naming including local knowledge and sensory ethnobotany
 - Inclusion enhances epistemic justice
- ..while keeping sight of overdetermination risk
 - Political economy of crop science and agricultural development - erases cultural, biological, scientific and semantic diversity
 - Inclusion enhances social *in*justice
- Epistemic justice can feed social injustice..
 - When is exclusion from global data linkage circuits a bad thing?
 - Severe data governance challenge





Thank you for your attention, and funders and collaborators for their support! This material was published as Leonelli, S. “Process-Sensitive Naming”, PTPBio 2022, <https://doi.org/10.3998/ptpbio.16039257.000000>

