Introduction to Ontology Semantics & tools for data annotation

Harold Duruflé

SUNRISE

Introduction to Ontology
Semantics & tools for data annotation

Harold Duruflé

Problem: Several words for retrieving the same data

(SUNRISE project)

Sunflower

Helianthus spp.

***Helianthus annuus* L.**

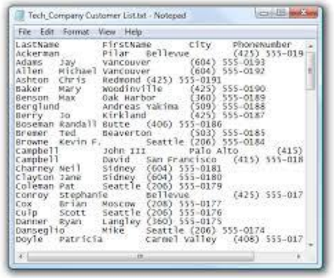
Tournesol

Helianthus

→ **Semantic heterogeneity (concepts)**

Problem: Data is multi-format and everywhere

Databases


Tech_Company Customer List.txt - Notepad
File Edit Format View Help

```
LastName        FirstName       City     PhoneNumber
Ackerman        Pilar   Bellevue         (425) 555-019
Adams   Jay     Vancouver        (604) 555-0193
Allen   Michael Vancouver        (604) 555-0192
Ashton  Chris   Redmond (425) 555-0191
Baker   Mary    Woodinville      (425) 555-0190
Benson  Max     Oak Harbor       (360) 555-0189
Berglund        Andreas Yakima   (509) 555-0188
Berry   Jo      Kirkland         (425) 555-0187
Boseman Randall Butte   (406) 555-0186
Bremer  Ted     Beaverton        (503) 555-0185
Browne  Kevin F.        Seattle (206) 555-0184
Campbell        John III        Palo Alto       (415)
Campbell        David   San Francisco   (415) 555-018
Charney Neil    Sidney  (604) 555-0181
Clayton Jane    Sidney  (604) 555-0180
Coleman Pat     Seattle (206) 555-0179
Conroy  Stephanie       Bellevue        (425) 555-017
Cox     Brian   Moscow  (208) 555-0177
Culp    Scott   Seattle (206) 555-0176
Danner  Ryan    Langley (360) 555-0175
Danseglio       Mike    Seattle (206) 555-0174
Doyle   Patricia        Carmel Valley   (408) 555-017
```

Tabulated files

text files

Publication / Reports

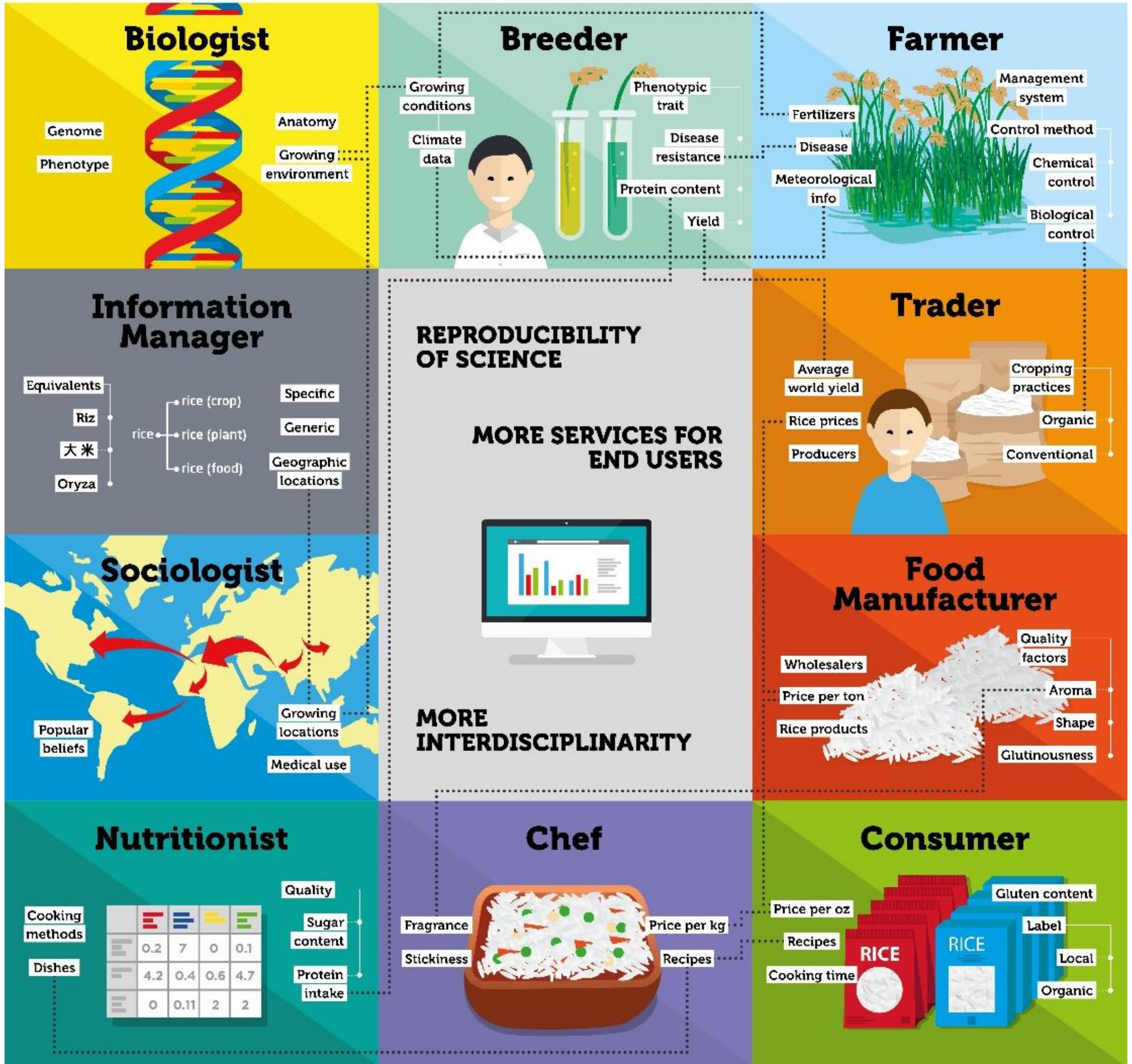
Excel files

→ Structural and syntactic heterogeneity

Several domains

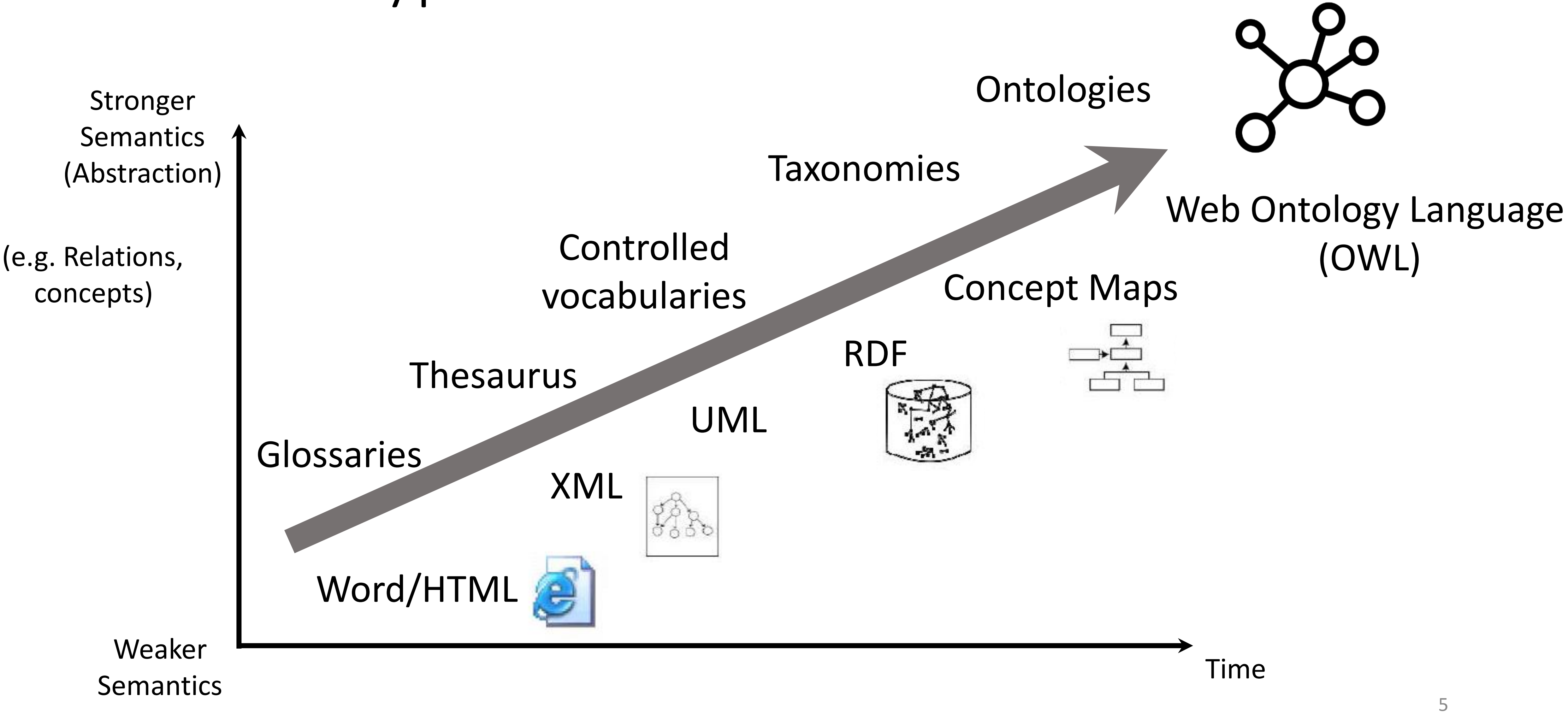
→ **Domain heterogeneity**

SEMANTICS - THE WAY TO RECONCILE POINTS OF VIEW AND DATA
THE EXAMPLE OF "RICE"



(INRA with the collaboration of the RDA Agrisemantics WG
http://aims.fao.org/activity/newsletter/aims-newsletter-no80-january-2019)

Different types of semantic resources



Stronger Semantics (Abstraction)

(e.g. Relations, concepts)

Weaker Semantics

Time

Word/HTML

Glossaries

XML

Thesaurus

UML

Controlled vocabularies

RDF

Taxonomies

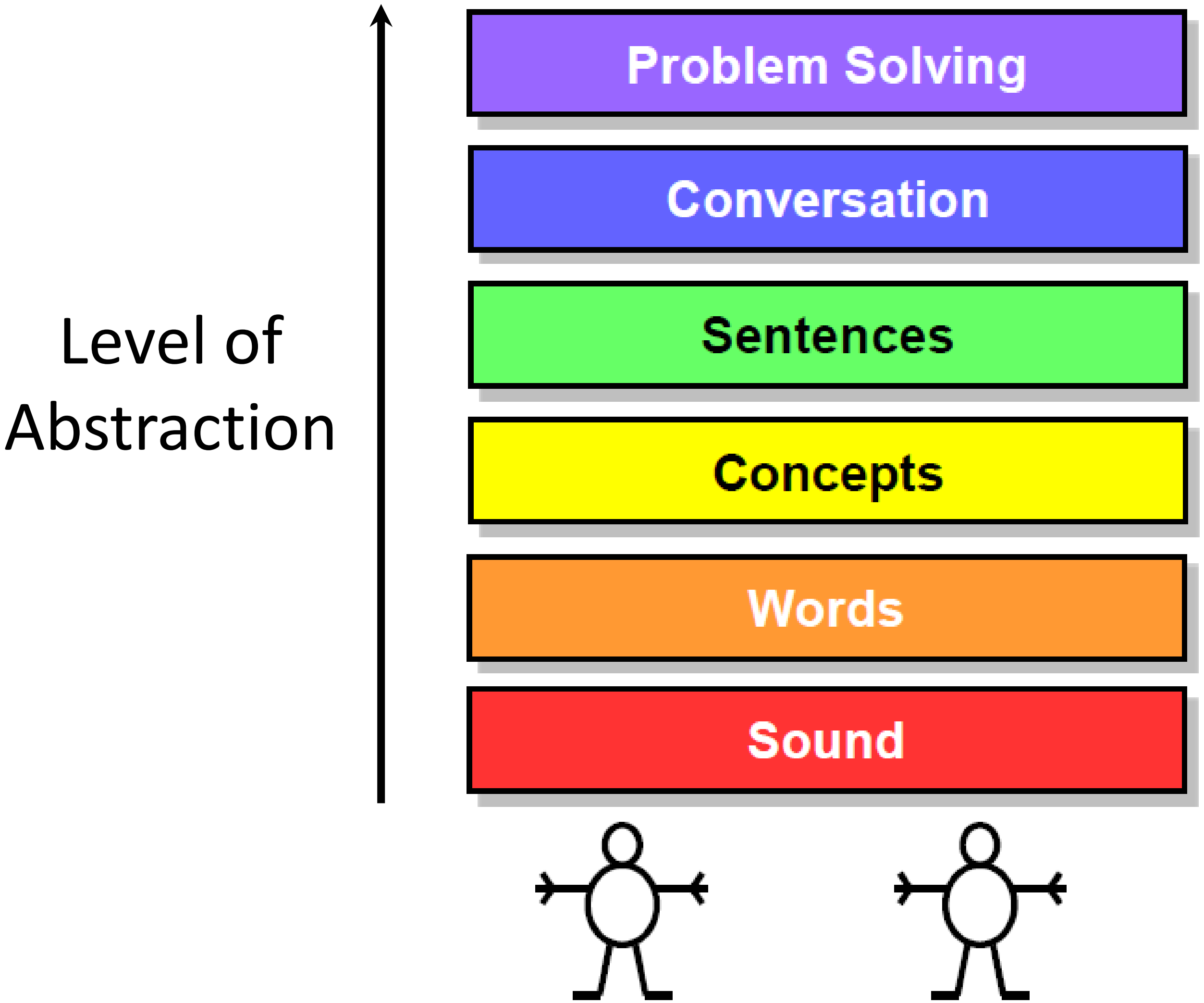
Concept Maps

Ontologies

Web Ontology Language (OWL)

(Adapted to Semantic Spectrum & Network Inference)

Example:



Level of Abstraction

Problem Solving

Conversation

Sentences

Concepts

Words

Sound

Yield Improvement

Pathways

Structural motif

Protein

Sequence

DNA

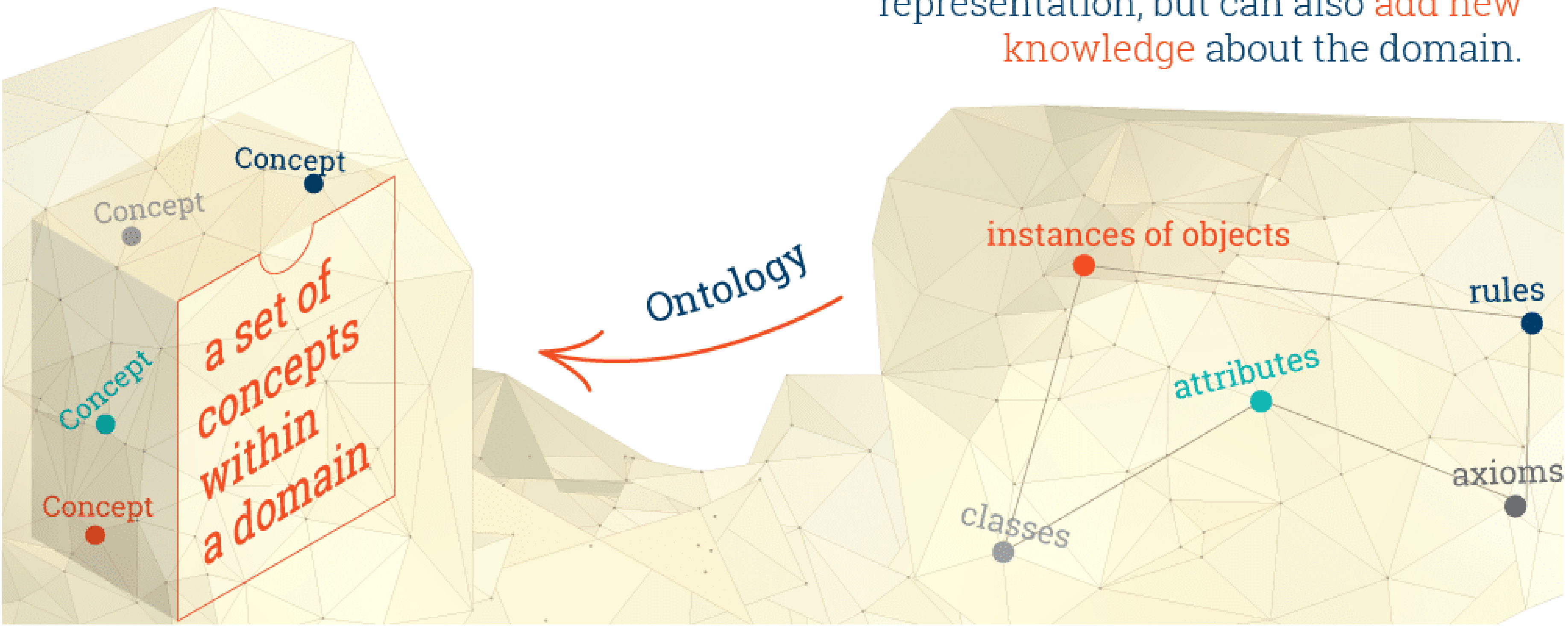
What is an ontology?

- Provide a shared vocabulary for a domain (all the terms)
- Provide textual definitions that describe the intended meaning of the terms in vocabularies
- Provide standard identifiers for concepts describing a given domain
- Provide machine-readable axioms and definitions that enable computational access to some aspects of the meaning of classes and relations – logical representation of human knowledge

→ **Facilitate data publication / data access and analysis**

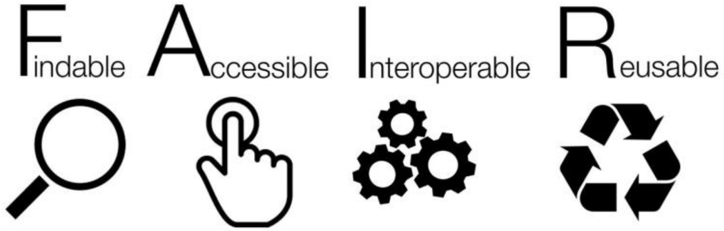
What is an ontology?

Ontologies do not only introduce a sharable and reusable knowledge representation, but can also add new knowledge about the domain.



https://www.ontotext.com/knowledgehub/fundamentals/what-are-ontologies/

FAIR principles



FAIR: Findable, Accessible, Interoperable, Reusable

(Wilikinson, 2016 nature DOI: DOI:10.1038/sdata.2016.18)

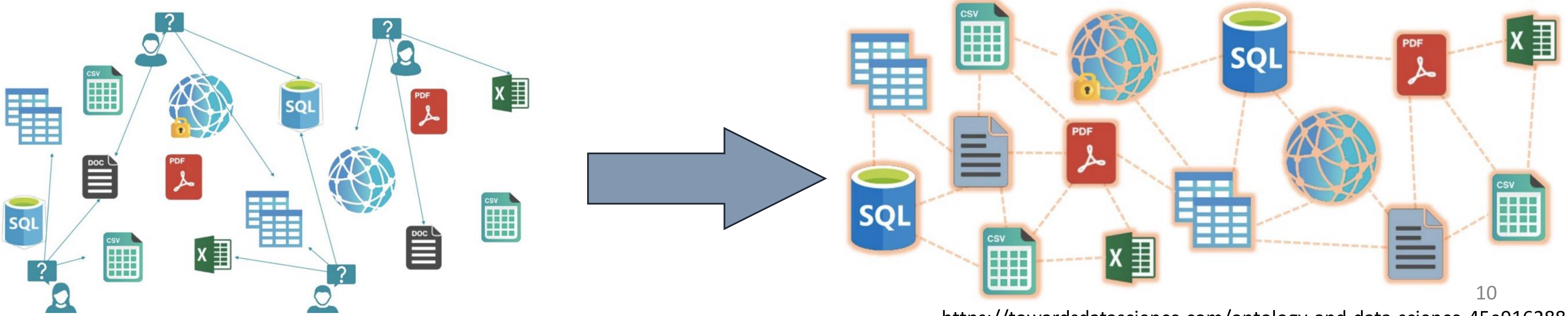
To be Interoperable:

- I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
- I2. (meta)data use vocabularies that follow FAIR principles.
- I3. (meta)data include qualified references to other (meta)data.

www.go-fair.org/fair-principles/

Why biologist have adopted ontologies?

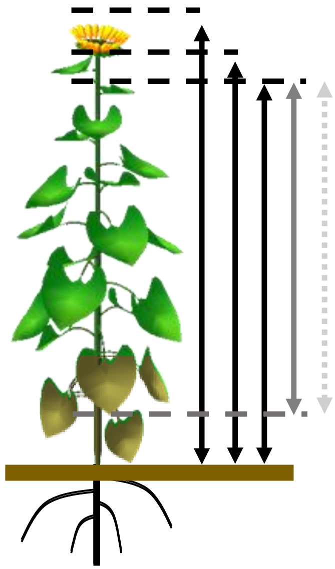
- To provide canonical representation of scientific knowledge
- To annotate experimental data to enable interpretation, comparison, and discovery across databases (Example: GO)
- To facilitate knowledge-based applications for
  - Decision support
  - Natural language-processing
  - Data integration



https://towardsdatascience.com/ontology-and-data-science-45e916288cc5

Example of the crop ontology

- No naming convention for variables and methods of measurement which are heterogeneous
- Trait & Variable definitions and measurement are not similar between farmers, breeders, agronomists, modelers,…



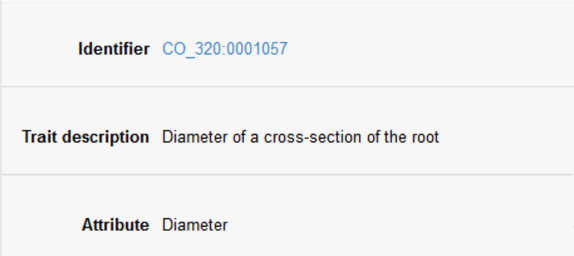
Plant Height

→ One trait = x traits…

Example of the crop ontology

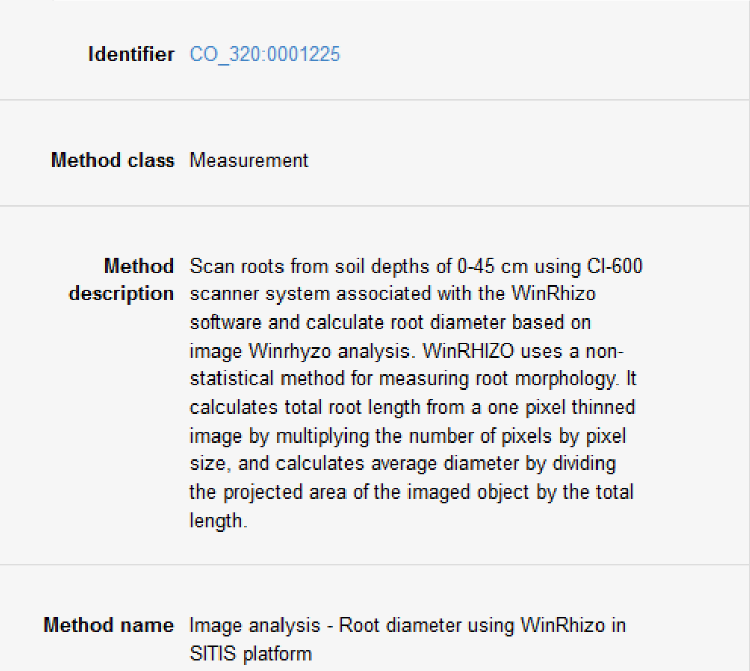
Annotation must explain:

1/ What is the observation about? = **TRAIT**
(e.g. Plant Height, Color of grain)



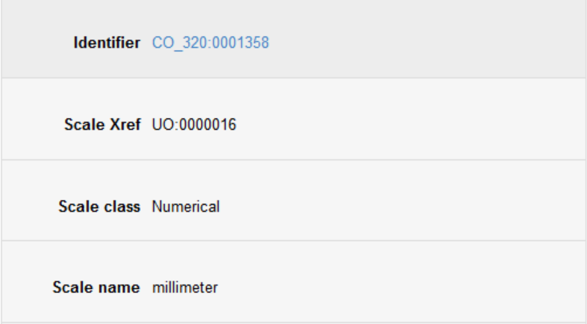
Identifier	CO_320:0001057
Trait description	Diameter of a cross-section of the root
Attribute	Diameter

2/ How is the trait observed? = **METHOD**
(e.g. Measuring, Estimated visually, Calculated)



Identifier	CO_320:0001225
Method class	Measurement
Method description	Scan roots from soil depths of 0-45 cm using CI-600 scanner system associated with the WinRhizo software and calculate root diameter based on image Winrhyzo analysis. WinRHIZO uses a non-statistical method for measuring root morphology. It calculates total root length from a one pixel thinned image by multiplying the number of pixels by pixel size, and calculates average diameter by dividing the projected area of the imaged object by the total length.
Method name	Image analysis - Root diameter using WinRhizo in SITIS platform

3/ How is the trait observation expressed? = **SCALE**
(e.g. cm, short/medium, white/black)



Identifier	CO_320:0001358
Scale Xref	UO:0000016
Scale class	Numerical
Scale name	millimeter

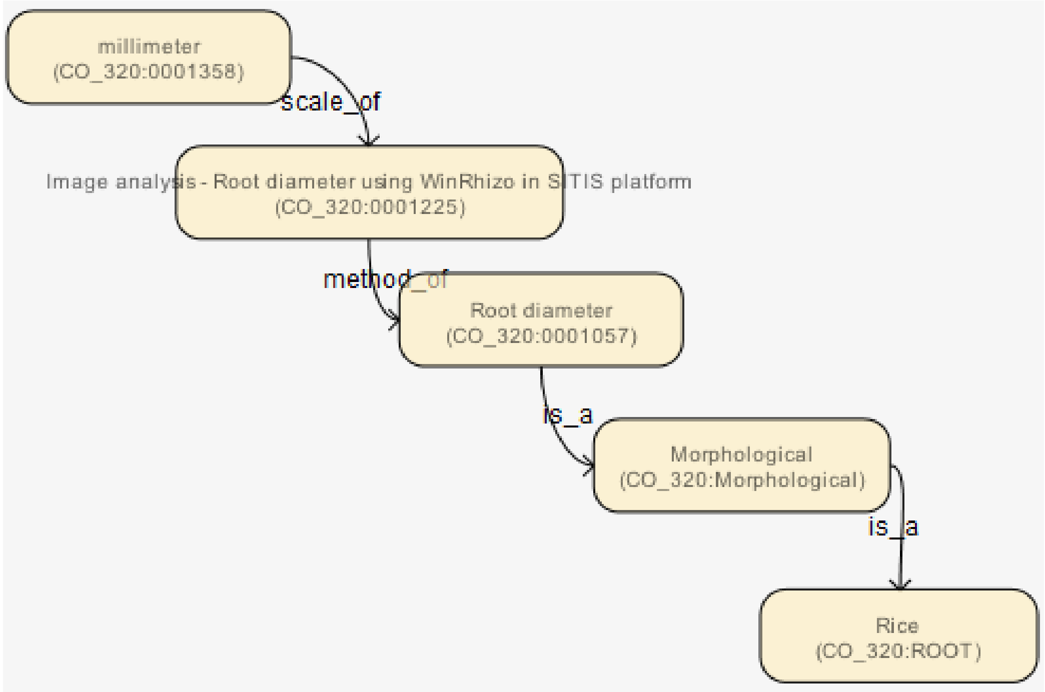
Example of the crop ontology

Annotation must explain:

1/ What is the observation about? = **TRAIT**
(e.g. Plant Height, Color of grain)

2/ How is the trait observed? = **METHOD**
(e.g. Measuring, Estimated visually, Calculated)

3/ How is the trait observation expressed? = **SCALE**
(e.g. cm, short/medium, white/black)



millimeter (CO_320:0001358)

scale_of

Image analysis - Root diameter using WinRhizo in SITIS platform (CO_320:0001225)

method_of

Root diameter (CO_320:0001057)

is_a

Morphological (CO_320:Morphological)

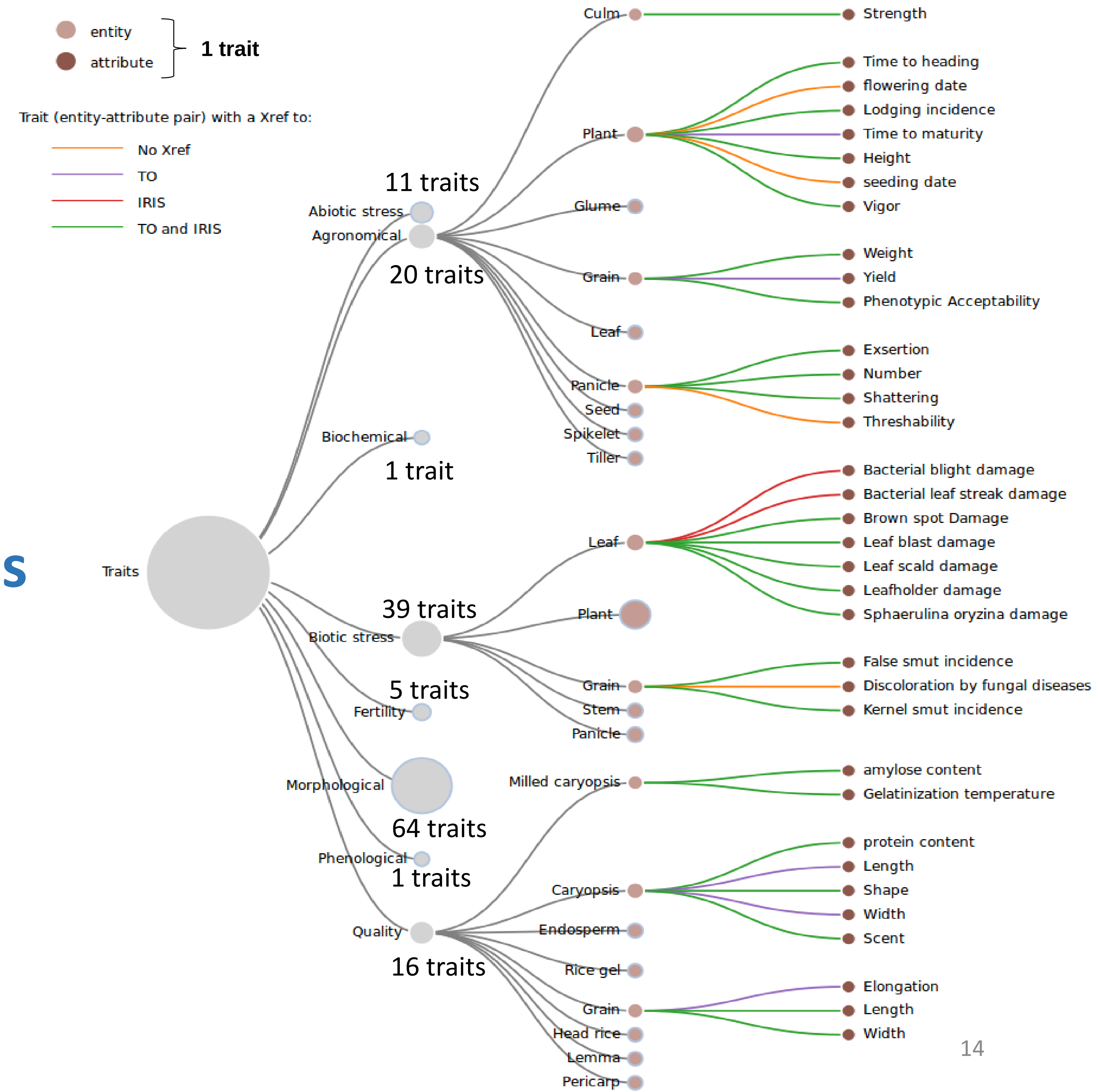
is_a

Rice (CO_320:ROOT)

Example of an ontology

Crop Ontology
for agricultural data

Rice Traits



Example of link between ontologies

fruit color trait (TO:0002617) across various species

Fruit Color

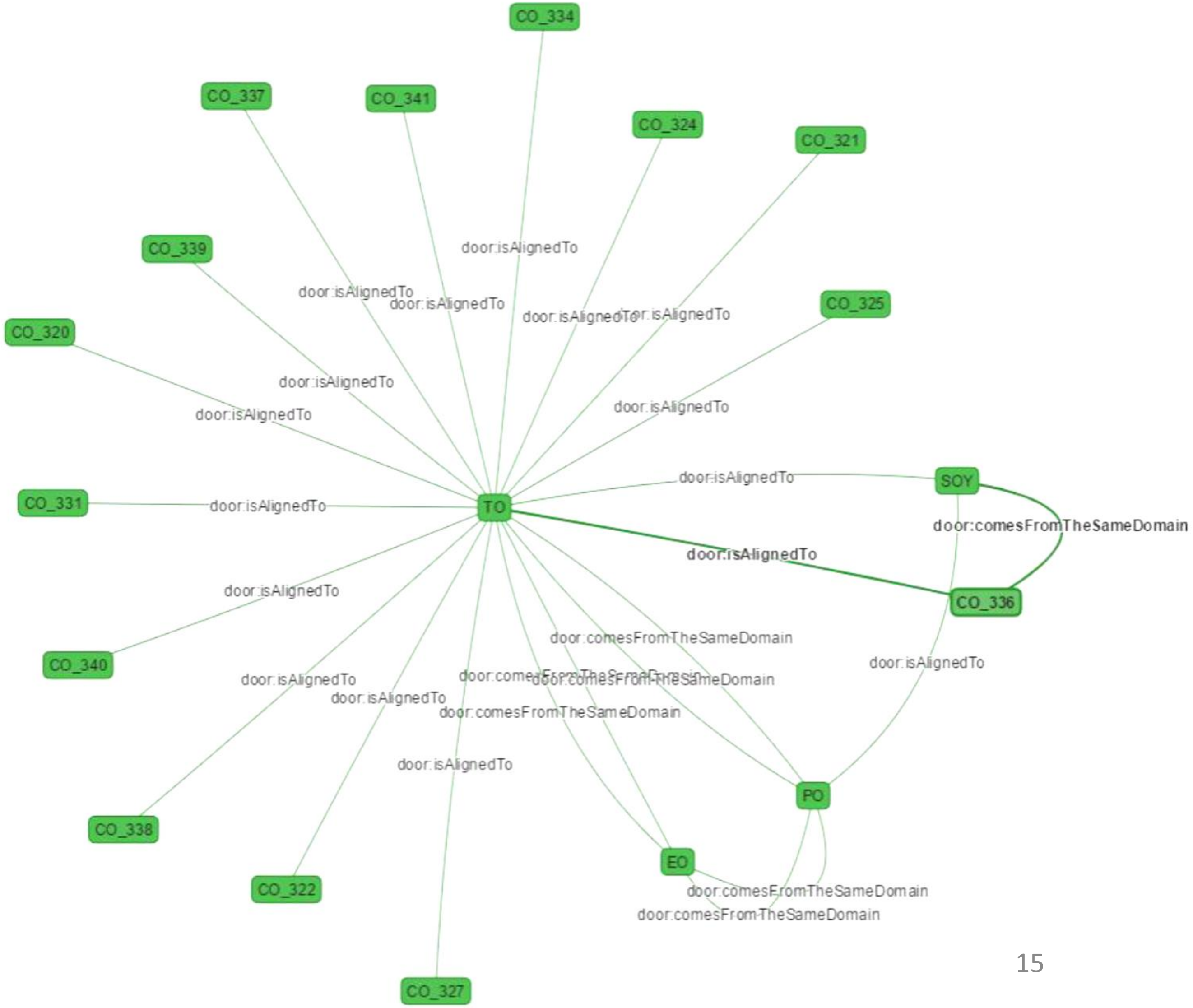
Kernel Color

Grain Color

Achene Color

Berry Color

Pod Color

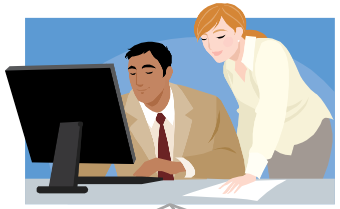


→ **Ontologies can link to data from multiple species**

Crop Ontology Workflow

Breeding Management System

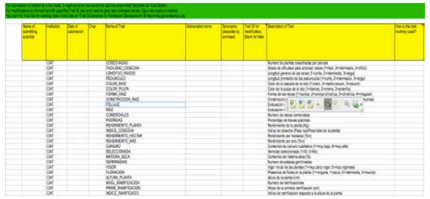
Breeders, Data Manager & Scientists



Plants' traits ontology

Defined and organized variables

In Trait Dictionary

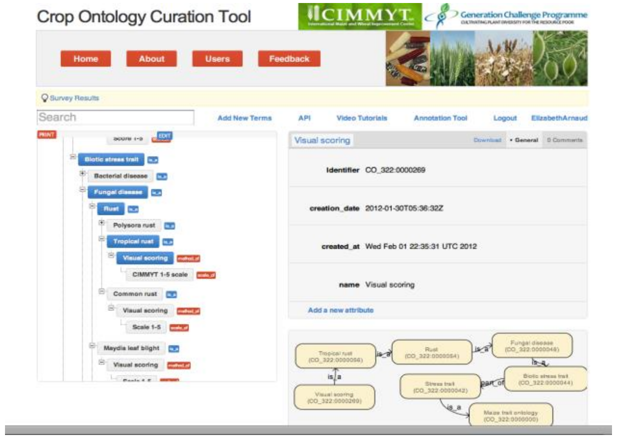


[illegible]

Name of submitting scientist	Institution	Date of submission	Crop	Name of Trait	Abbreviated name	Synonyms (separate by commas)	Trait ID for modification, Blank for New	Description of Trait	How is the trait routinely used?
	CIAT			COSECHADAS				Numero de plantas cosechadas por parcela	
	CIAT			FACILIDAD_COSECHA				Grado de dificultad para arrancar raices (1=facil, 2=intermedio, 3=dificil)	
	CIAT			LONGITUD_RAICES				Longitud general de las raices (1=corta, 2=intermedia, 3=larga)	
	CIAT			PEDUNCULO				Longitud promedio de los pedunculos (1=corto, 2=intermedio, 3=largo)	
	CIAT			COLOR_RAIZ				Color de la cascara de la raiz (1=claro, 2=medio oscuro, 3=oscuro)	
	CIAT			COLOR_PULPA				Color de la pulpa de la raiz (1=blanco, 2=crema, 3=amarillo)	
	CIAT			FORMA_RAIZ				Forma de las raices (1=conica, 2=conica-cilindrica, 3=cilindrica, 4=irregular)	
	CIAT			CONSTRICCION_RAIZ				Constriccion [illegible]	
	CIAT			FOLLAJE				Evaluacion [illegible]	
	CIAT			RAIZ				Evaluacion [illegible]	
	CIAT			COMERCIALES				Numero de raices comerciales	
	CIAT			PODRIDAS				Porcentaje de raices podridas	
	CIAT			RENDIMIENTO_PLANTA				Rendimiento de la planta (kg)	
	CIAT			INDICE_COSECHA				Indice de cosecha (Peso raiz/Peso total de la planta)	
	CIAT			RENDIMIENTO_HECTAR				Rendimiento por hectarea (Ton)	
	CIAT			RENDIMIENTO_ANO				Rendimiento por ano (Ton)	
	CIAT			CIANURO				Contenido de cianuro cualitativo (1=muy bajo, 9=muy alto)	
	CIAT			SELECCIONADA				Variedad seleccionada (1=Si, 0=No)	
	CIAT			MATERIA_SECA				Contenido de materia seca (%)	
	CIAT			GERMINADAS				Numero de estacas germinadas	
	CIAT			VIGOR				Vigor inicial de las plantas (1=muy poco vigor, 5=muy vigorosa)	
	CIAT			FLORACION				Presencia de flores en la planta (1=ninguna, 1=poca, 2=intermedia, 3=mucha)	
	CIAT			ALTURA_PLANTA				Altura de la planta (cm)	
	CIAT			NIVEL_RAMIFICACION				Numero de ramificaciones	
	CIAT			PRIMR_RAMIFICACION				Altura de la primera ramificacion (cm)	
	CIAT			INDICE_RAMIFICACIO				Indice de ramificacion respecto a la altura de la planta	

Crop Ontology

for agricultural data



Crop Ontology Curation Tool

Identifier CO_322:0000269

creation_date 2012-01-30T05:36:32Z

created_at Wed Feb 01 22:35:31 UTC 2012

name Visual scoring

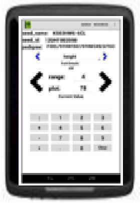
Published online

Possibility to download variables

Possibility to upload variables

Fieldbook creation

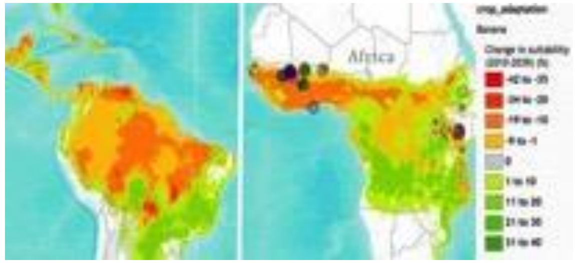
Ease data collection in the field thanks to defined variables



Database

Store annotated data, and allow their interpretation and harmonization

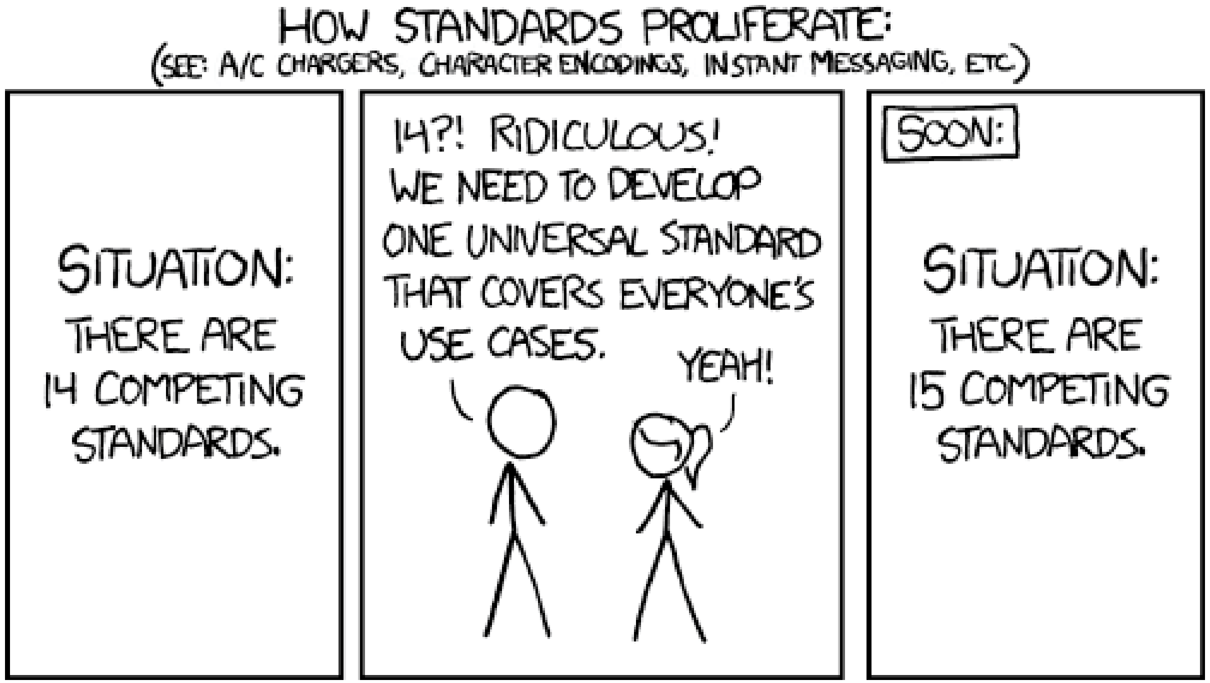
Data Analysis



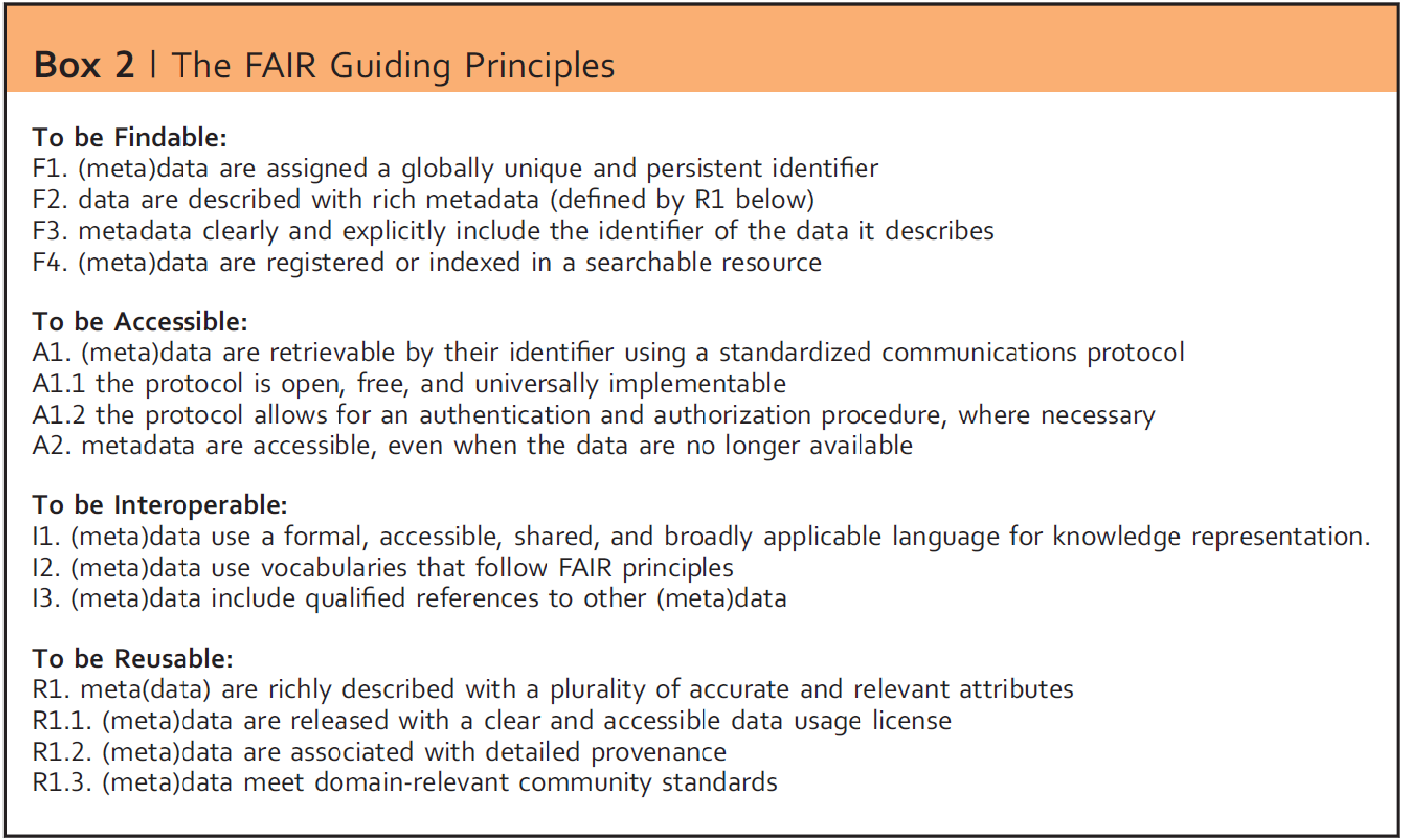
Thank you

- Bioversity international CGIAR
  - **Elizabeth Arnaud**
  - **Marie Angélique Laporte**

- **Cyril Pommier**


HOW STANDARDS PROLIFERATE:
(SEE: A/C CHARGERS, CHARACTER ENCODINGS, INSTANT MESSAGING, ETC)
SITUATION: THERE ARE 14 COMPETING STANDARDS.
14?! RIDICULOUS! WE NEED TO DEVELOP ONE UNIVERSAL STANDARD THAT COVERS EVERYONE'S USE CASES.
YEAH!
SOON:
SITUATION: THERE ARE 15 COMPETING STANDARDS.

(xkcd.com/927)



Box 2 | The FAIR Guiding Principles

To be Findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. data are described with rich metadata (defined by R1 below)
F3. metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource

To be Accessible:
A1. (meta)data are retrievable by their identifier using a standardized communications protocol
A1.1 the protocol is open, free, and universally implementable
A1.2 the protocol allows for an authentication and authorization procedure, where necessary
A2. metadata are accessible, even when the data are no longer available

To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation.
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data

To be Reusable:
R1. meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

(Wilikinson, 2016 nature DOI: DOI:10.1038/sdata.2016.18)