Crop Ontology
for agricultural data

(photo credit: Carlo Fadda)

Guidelines to create a Crop Ontology for phenotype annotations

Léo Valette, Bioversity International

Julian Pietragalla, Integrated Breeding Platform

Bioversity International

Version: 11 May 2018

Introduction

Crop Ontology (CO, http://www.cropontology.org) is a multi-purpose resource of the **Integrated Breeding Platform** (IBP, http://integratedbreeding.net/).

CO is an infrastructure that centralizes and serves the **semantics** used by the breeders to the **Breeding Management System** and to third party information systems through an API (Application programming interface). The semantics allows annotating germplasms, phenotypes, plant anatomical structures and developmental stages as well as the environmental and experimental conditions of the trial. Besides the infrastructure, CO is a **standard framework** to build ontologies that annotate crop-specific phenotypes.

These ontologies follow a **conceptual model** where the combination of a trait, a method and a scale determines a phenotypic observation variable. This model aims at supporting the edition of breeders' fieldbooks and the data annotation while making the annotated data **discoverable**, **retrievable** and **interoperable**.

The framework relies on a workflow designed for the community. A crop-specific community creates a **Trait Dictionary** (TD) by formatting its phenotypic observation variables in a **template**. The TD is the support to submit, curate and harmonize the variables. Once the community has finalized the TD, it is published on cropontology.org (http://www.cropontology.org/add-ontology). This step implements the conceptual model and generates unique identifiers.

This document details the CO model and explains how to practically create a TD.

The CO phenotype annotation model

The CO concepts

Variable

Breeders Record Variables

The CO phenotype annotation model is grounded in breeder's datasets. The column headers generally integrate a lot of codified information. By way of example, Table 1 shows a breeder's dataset subset.

Table 1: example of a subset of a breeder's dataset

Germplasm ID	PH	GCOL	GY	…
24530	80	2	35	
85432	120	4	24	
78452	90	4	30	
…				

Reading the project report is essential to understand that "PH", "GCOL" and "GY" are breeders' terms that stand for, respectively:

- PH: “the height of the plant, defined as the distance from the ground to the top of the canopy, that is measured with a ruler and expressed in cm”
- GCOL: “the color of the grain visually assessed expressed on a codified categorical scale where 2=yellow and 4=orange”
- GY: “grain yield which is derived by dividing the weight of dehulled grains harvested from the plot by the surface of the plot and which is reported in g per m²”

CO calls “*PH*”, “*GCOL*” and “*GY*” **variables**. The CO model standardizes the definition of breeders’ variables. The purpose of this standardization is two-fold:

1. The model provides breeders with standard variables that can be readily used to **edit breeding fieldbooks**
2. The model makes **data integration** possible by formalizing the different pieces of information that compose a variable

1 variable = {1 trait, 1 method, 1 scale}

The essence of the CO model is to decompose a variable recorded by the breeders into:

- A **trait**: “what is observed”
- A **method**: “how the observation is made”
- A **scale**: “how the observation is expressed“

In other words, **a variable is the combination of 1 trait, 1 method and 1 scale**. To illustrate this, the Table 2 shows the breakdown of the PH, GCOL and GY examples.

Table 2: Breakdown of PH, GCOL and GY into trait, method and scale

Variable	Trait	Method	Scale
PH	The distance from the ground to the top of the canopy	Measured with a ruler	cm
GCOL	Color of the grain	Visually assessed	5-category color scale
GY	Yield of dehulled grain	Divide harvested grain weight by plot surface	g/m^2

Breaking down the variables into trait, method, scale helps data integration. It allows querying heterogeneous databases such as datasets of breeding trials, gene markers, agronomy trials, and retrieve the data points that share the same trait and/or method and/or scale. By way of example, databases with CO annotated data can be queried to retrieve the data points that refer to the height of the plant, regardless of the observation method (direct measurement, visual assessment, mean computation of direct measurements, etc…) and regardless of the reporting unit or the scoring system.

Trait

The trait is a combination of entities and attributes

CO defines a trait as **a character of an individual plant or of a group of plants that can be observed and that results from the expression of its genes and their interaction with the environment**. This trait definition can be summarized by "what is observed". Examples of traits are "plant height', "pod color", "grain yield", "seed germination rate", "plant flowering time", "panicle shape", "plant resistance to blight", etc.

Traits can be formalized by a **meaningful combination** of **entities** and **attributes**. The entity is the **observed plant part**. The entity can be defined at the level of the cell, the tissue, the organ, the whole organism or the sub-product of the crop. The attribute is the **feature of an entity**. Attributes can be states or processes. Examples of states are weight, length, area, color, chemical content, while examples of processes are vegetative period duration, photosynthesis rate, senescence rate.

A direct consequence of the trait decomposition into entity and attribute is that two traits that share the same combination of attributes are identical even though they are named differently. Thus, decomposing a trait into entity/attribute is a good practice to prevent trait duplication. Table 3 shows examples of trait breakdown into entity/attribute.

A trait can be described by a general entity or as a specimen, something that obviously belongs to an entity but is noticed by reason of an individual distinguishing characteristic. Examples of specimen are primary branch, first flower, first grade commercial fruit, flag leaf, third leaf, main stem, fertile stem, mature fruit, immature fruit.

Table 3: Examples of trait breakdown into entity and attribute

Trait	Entity	Attribute
Plant height	Plant	Height
Number of secondary branches	Secondary branch	Number
Plant flowering time	Plant	Flowering time
Plant phenotypic acceptability	Plant	Phenotypic acceptability
Leaf chlorophyll content	Leaf	Chlorophyll content
Leaf area index	Leaf	Area index
Plant rust severity	Plant	Rust severity
Flour gluten content	Flour (milled grain)	Gluten content
Dough elasticity	Dough	Elasticity

Traits are grouped in consensus classes

The 2014 Crop Ontology workshop led to a set of 9 consensus grouping classes. Table 4 inventories the grouping classes and provides a definition.

Table 4: Definition of the trait grouping classes

Trait class	Definition	Examples of traits
Morphological	All traits related to anatomical (internal) and morphological (external) structure of the plant, its organs and tissues.	- Fruit shape - Seed color - Stem diameter - Seed length
Phenological	All traits related to growth/developmental stages and periods of crop/plants.	- Plant flowering time - Plant maturity time - Plant vegetative period duration
Physiological	All traits related to functioning of the crop/plant and its response/adaptation to the environment.	- Leaf senescence rate - Canopy temperature - Canopy NDVI - Leaf stomatal conductance
Agronomical	All main traits contributing to yield and related to the agronomic performance of crop/plants.	- Seed/tuber yield - Biomass yield - Plant lodging incidence
Biotic stress	All traits related to stress caused by living stressors. Biotic stress is defined as the negative impact on the crop/plants of living organisms such as bacteria, viruses, fungi, parasites, nematodes, weeds, invertebrate and vertebrate pests.	- Plant disease severity - Plant disease incidence - Plant pest damage - Plant disease plant response
Abiotic stress	All traits related to stress caused by non-living stressors. Abiotic stress is defined as the negative impact of non-living factors on the crop/plants. Most common abiotic stressors are drought, waterlogging, high/low temperatures, mineral toxicities/deficiencies, hail, and wind.	- Plant aluminum tolerance - Plant drought susceptibility - Plant frost damage - Plant heat tolerance
Biochemical	All traits related to chemical components of a plat entity.	- Leaf ABA content - Seed Proline content - Tuber carotenoid content
Quality	All traits related to key characteristics that influence end-use quality of crop/plant products (seed, fruit, leaf, root/tuber, etc.) and sub-	- Seed protein content - Fruit sugar content

	products (flour, dough, pulp, etc.)	- Dough color - Pasta Consumer acceptability
Fertility	DEF	

There are cases where a trait can be classified in two or more classes, as far as possible select the most representative class.

Method

Definition

The method describes how the trait is observed which covers two notions: the sampling and the protocol or technics. Always try to keep the method as general as possible, for a detailed procedure add its reference publication.

Sampling

The **sampling** specifies whether the observation the plant entities that must be observed either one individual plant entity or on a collection of plant entities. However, if the sample is a set of more than one entity, the method should mention a recommended range, for example observe plant high in 5 to 9 plants and record the average.

Observation procedures

The method must also detail the **procedure** to follow in order to observe the sampled entity/entities. Examples of protocols are:

- Measurement of the plant height using a ruler. It is recommended to measure 5 to 10 plant and record the average.
- Leaf area derived from image analysis. Lay the leaf flat and take a picture with the lens set parallel to the leaf. Single out the pixels of the leaf by filtering the image for contrast using <software>. Count the number of leaf pixels and multiply the count by the area represented by each pixel.
- Visual estimation of the fruit colour.
- Measure grain weight with a weighing scale.
- Visual assessment of the leaf color based on a standard color chart.

It is important to clarify here that the observation procedure is different than the experimental protocol. The experimental protocol combines the trial design, treatment factors and the experimental conditions. The observation procedure focuses on strictly defining how the observation is made, not how the trial is led.

The method classes

Table 5: Method class definitions

Method class	Type	Definition	Examples of methods
Measurement	Direct observation	The trait observation is supported by a measuring device, a sensor.	- Measurement with a ruler

			- Weighing on a scale
Counting		The trait is observed by counting entities.	- Leaf counting
Estimation		The trait is directly assessed without the help of a measuring device. Estimations considerably rely on the subjectivity of the observer.	- Grain colour estimation - Damage on leaves visual estimation - Stem height visual estimation
Computation	Indirect observation	A computation is a method that indirectly observes a trait by computing more than one direct observation.	- 100 grain calculation (100 * measured grain weight / grain count) - Obtain the seed protein to seed oil content ratio

Scale

Definition

The scale describes how the trait observation is **expressed**.

The scale classes

Scales are grouped in scale classes. Their descriptions are summarized in Table 6.

Table 6: Scale classes

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 exemple a severity trait might be expressed by a 2 digit and 2 character code. The first 2 digits are the percentage of the plant covered by a fungus and the 2 characters refer to the delay in development, e.g. “75VD” means “75% of the plant is infected and the plant is very delayed”.
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 and fixed 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 scales composed of ordered categories.
Text	A free text is used to express the trait.

Date	The date class is for events expressed in a time format, e.g. "yyyymmdd hh:mm:ss – UTC" or "dd/mm/yy"

It is mandatory to applied (reutilize) existing scales id. Note: there are crops ontologies published where different scales ids were applied to the same scale, this situation is being regularized by Marie-Angelique Laporte from Crop Ontology team.

The limits and workaround of the CO model

The prerequisite of the CO model is that ontologies can be developed and maintained by the breeding community. Consequently, the data model has been kept as simple as possible.

However the CO model does not fulfil all the community needs for phenotype data annotation and integration. This section covers some of the gaps and shows some workarounds.

Time series and subsampling management

CO has not implemented a solution to handle repeated observations of the same variable in time and in space. While an harmonized solution is not developed, these aspects may be handled by adding components to the variable name string.

Time series observations

To monitor the growth of the plant or an outbreak of a disease for example, breeders repeat several observations of the same trait, with the same method and express the observation in the same scale. In these cases, the same CO variable is observed at different times. Breeders have to make the difference between these observations in their fieldbooks and databases.

Though this issue has been clearly identified, the CO model does not offer a solution to leverage this issue. The main obstacle is that the solution would have to be flexible enough to allow any type of time stamping: date (e.g. yyyymmdd), phenological stage (e.g. at flowering), duration from a phenological stage (e.g. 1 month after sowing).

For now, repeated observations of the same CO variables have to be managed at the level of the project databases.

ATTENTION It does not mean that growth stage is not part of the trait – only if the growth stage implies observing something different (trait), observing differently (method)

Observations on subsampled entities

In the context of high throughput phenotyping or advanced physiology experiments particularly, scientists observe the same variable on several equivalent plant entities. A dummy example is the assessment of the colour of the leaf on each single leaf of each plant. In this example, CO considers that there is only one entity that is observed, the leaf. The CO model cannot generate a variable for each leaf that can potentially be observed. Yet, the project fieldbooks and databases have to differentiate all these measurements.

Temporary solution offered by the BMS

The IBP is testing an additional template to manage repeated observations of the same CO variable. The principle is that for each of CO variable with time series measurement, a time stamp is defined and given a code. The system then processes the template and generates

local BMS variables by creating a local unique identifier and appending the time stamp code to the CO variable name: <CO variable name>_<time stamp code>.

A similar process is applied for CO variables that are repeatedly observed on equivalent plant entities resulting in a local BMS variable named as follow: <CO variable name>_<time stamp code>_<subsampled entity code>

For further information, please contact Julian Pietragalla, IBP technical support, at j.pietragalla@cgiar.org

Methods can exceptionally refine the trait entity

Observation protocols, in some cases, include and describe processing operations such as grinding, milling, polishing, drying. These operations give detailed information regarding the entity that is observed e.g. fresh or dry plant, hulled or dehulled grain. Based on the definition of the CO "trait" concept, this information about the entity is to be defined in the trait description.

Yet, two variables that differ by the entity that is observed can exceptionally be differentiated at the level of the method. Let's take the two variables "fresh biomass in gram" and "dry plant biomass in g" as an example. The two decompositions into trait, method and scale that can be envisaged are shown in the table below.

	Trait-Method-scale decomposition	Variable 1: "plant fresh biomass in gram"	Variable 2: "dry plant biomass in gram"
Option A: 2 traits, 1 method	Trait	Plant fresh biomass	Plant dry biomass
	(2 different trait entities)	(Fresh plant)	(Dry plant)
	Method	Weight measurement	Weight measurement
	Scale	Gram	Gram
Option B: 1 trait, 2 methods	Trait	Plant biomass	Plant biomass
	(1 trait entity)	(Plant)	(Plant)
	Method	Fresh weight measurement	Dry weight measurement
	Scale	Gram	Gram

Option A is how the CO model recommends proceeding. A crop data manager who wishes to group those two variables sometimes prefers option B. The CO can accept such exceptions on account of this reason. However, the data managers must be informed that the Planteome project [LINK/REF] aims at mapping the crop specific ontology terms to the high level terms of the Trait Ontology [LINK]. This initiative will offer a hierarchy that will group the related terms of each crop-specific ontology. [For more information, contact m.a.laporte@cgiar.org]

The granularity of the terms

The granularity of the terms defines how specific the term is. The more specific the term is, the better it describes the actual measurement but the more it limits data integration.

Though CO intends to create ontologies applied to the plant breeding domain, there is no rule as to how specific the terms must be. There is a trade-off to be found based on this question: "*what is the impact of adding/removing a piece of information in the definition of the term on the interpretation of the data?*"

The Trait Dictionary

The CO model is implemented by uploading a Trait Dictionary (TD) on the Crop Ontology platform (http://www.cropontology.org/add-ontology). The TD is a flat but has a structured excel template.

The TD structure

In essence, the TD template is a two-dimension table. The columns represent the properties of the variable, trait, method and scale concepts. The rows represent the instances of variables and their corresponding trait, method and scale (see Figure 1)Figure 1.

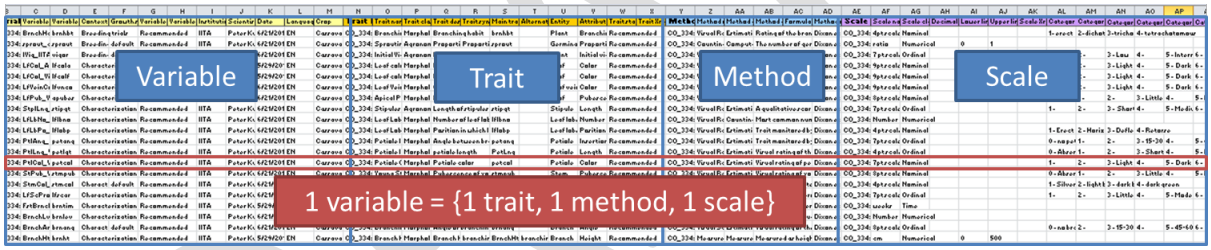

Figure 1: Structure of the TD template

Within a TD, each variable appears only once. However, a given trait, method or scale can be part of more than one variable. Terms composing more than one variable have to be **identically replicated** in the TD (see example in Figure 2).

B	C	N	O
Variable ID	Variable name	**Trait ID**	Trait name
CO_320:0000671	CaryoLng_Av_mm	CO_320:0000110	caryopsis length
CO_320:0000672	CaryoLng_MeasSES_	CO_320:0000110	caryopsis length

Figure 2: Example of the trait "caryopsis length". It is duplicated in the TD because it is part of the "CaryoLng_Av_mm" and "CaryoLng_MeasSES_1to7" variables

The CO identifiers

CO associates a unique identifier (ID) to each ontology term. IDs are constructed by concatenating a crop code identifier and a 7-digit numerical identifier (see Figure 3). The list of the already attributed crop codes can be found in Table 7 of the Appendix.

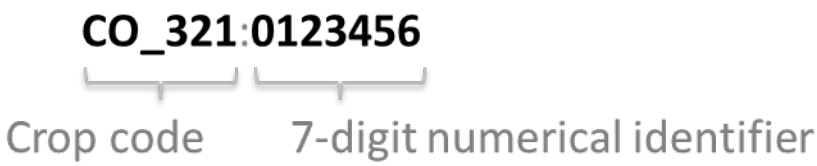

Figure 3: Structure of a CO identifier