Crop Ontology Governance and Stewardship Framework

Contents

This proposal emanates from a workshop held with most of the curators and some contributing scientists, key partners to define the framework (see the 29 participants in annex).

OBJECTIVES

A governance & stewardship framework for the Crop Ontology Project is required as this is a collaborative tool developed by a Community of Practice. Over the last 12 years of its existence, it has increased significantly in scope and use. Collecting and storing plant trait data and annotating the data with ontology terms is widely accepted by the crop science community to be critical to enable data interoperability and interexchange through tools such as the Breeding API (BrAPI). The Crop Ontology Community of Practice is organised around roles, curation principles and validation processes that require a formal description. A governance framework is defined by the various actors involved in the asset's design, development and maintenance. It is complemented by a quality assurance process to ensure that trust levels, value creation, and sustainability objectives meet appropriate quality levels. The general principles underlying data governance are integrity, transparency, accountability and ownership, stewardship, standardization, change management and a robust data audit.

BACKGROUND

The Crop Ontology Project (CO) was established by the CGIAR in 2009 and is maintained thanks to CGIAR funds (Integrated Breeding Platform and AgriFood Systems Research Programmes) as well as a 4-year grant from US NSF #1340112- Planteome project. The CO is a collaborative public good, released with the CC-By 4.0 license. It is currently made up of 33 crop ontologies developed by the 15 CGIAR centers and around 10 Non-CGIAR institutions, universities, and industry groups.

> Crop Ontology in January 2022
>
> - 33 species
> - 5,444 Traits
> - 7,507 variables
> - Contributors
>   - 15 CGIAR current centers
>   - Over 10 contributing Non-CGIAR institutions

For over 12 years, the global management and technical maintenance of the CO has been provided by the Alliance Bioversity-CIAT. Each crop-specific ontology is developed and maintained by a designated curator from a lead center or CGIAR institution for which the crop is mandated. The curator is often supported by a group of experts and interacts with a wde group of contributors and end-users from academia, R&D in agriculture, and the agri-food industry.

Eleven of the CO crops have been is successfully mapped to the Plant Trait Ontology (TO, https://obofoundry.org/ontology/to.html) of the Planteome Project (https://planteome.org/) to provide a path to search for traits across species, and support data retrieval for comparative genomics, amongst other use cases. The CO is also recommended by a major metadata schema for plant phenotype experiments called Minimum Information About a Plant Phenotyping Experiment (MIAPPE; https://www.miappe.org/) and by the standard Breeding API (BrAPI; https://brapi.org/) that is applied to breeding databases for the exchange of genotypic and phenotypic data. CO is included into the core database models of the Breeding Management System of the Integrated Breeding Programme, as well as of the Breedbases system developed by BTI. It is also used by platforms such as GEMS (https://agroinformatics.org/; Silverstein K., 2019). The CO is indexed and replicated by several ontology registries, for instance the Ontology LookUp Service (OLS;https://www.ebi.ac.uk/ols/) of the EMBL-EBI and by Agroportal (http://agroportal.lirmm.fr/), as well as in data standards registries like ELIXIR Fairsharing (https://fairsharing.org/).

If the CO does not fully comply with the OBO-Foundry Principles because of its structure (i.e. Principle 2: CO compiles crop specific trait ontologies with a primary focus on breeding and their domain knowledge partially overlaps with the Trait Ontology), it however complies with several of the quality criteria identified by the Ontologies Community of Practice (Arnaud et al., 2020). Table 1 shows the criteria that an ontology must address.

Table 1: Criteria established by CoP experts for identifying quality ontologies for data annotation. In orange are criteria requiring a Governance framework. Criteria in white are not addressed by CO (from Arnaud et al., 2020).

	Classified by the Panel of Experts
1	Adhere to the Open Biomedical Ontology (OBO) Foundry guidelines on design and format
2	Represent a unique non-overlapping knowledge domain
3	Willingness to express and integrate multiple, evidence-based classification systems in the chosen domain
4	Logically-structured and defined
5	May contain relationships and dependencies on other reference ontologies
6	Represent accurate science supported by evidence
7	Open source and Creative Commons CC-BY or CC-0 license (https://creativecommons.org/)
8	Must be widely used in annotation and data capture
9	Support both inter- and intra-specific needs with species agnostic (core) and specific (extensions) resources that work together.
10	Sustainable funding sources
11	Human resources to manage (i.e. curators, editors, and developers)
12	Established ontology management system including roles and responsibility
13	Must be designed to answer both the computing and community needs
14	Must explicitly identify the communities of reference
15	Centralized maintenance of the validated content, and distributed contribution and access
16	Quality assurance on the ontology by experts in the field of knowledge
17	Reducing reliance on internal processes and data stewardship networks

ELEMENTS OF THE GOVERNANCE FRAMEWORK

ONTOLOGY DEVELOPMENT AND MAINTENANCE: ROLES AND RESPONSIBILITIES

The development and maintenance of crop ontologies involves the actions and interactions in and across various domains. Figure 1 shows a diagram of the typical workflow for the creation of a Trait Dictionary that will be published online in the Crop Ontology.

There are three domains which interact:

1) Crop-specific domains (Orange)

2) CO global project domain (Blue)

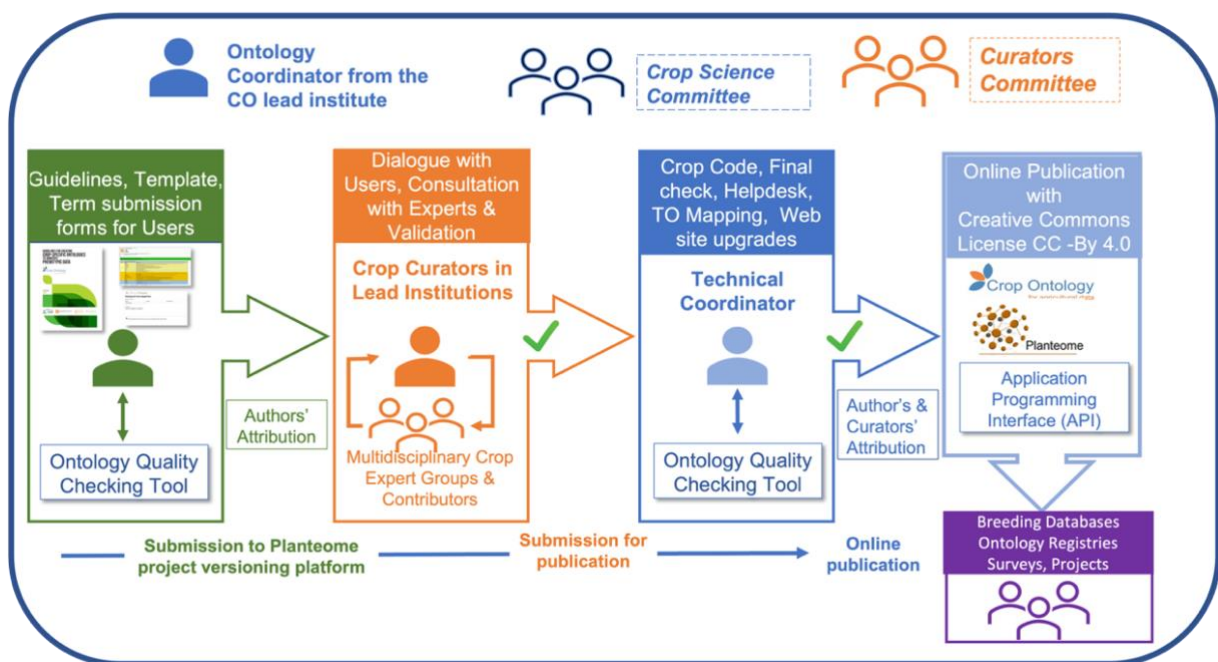3) External domain (Purple color i.e. breeding DB, Ontology Registries etc)

Figure 1: Roles & Workflow for the creation of a Crop-specific ontology and submission of terms

ROLES AT THE CROP-SPECIFIC DOMAIN LEVEL

Actors	Responsibilities
Crop science community	● Formulate the need for a specific Trait Dictionary to support their crop-related work ● The community identifies a potential lead institute capable to lead such an effort ● Provide feedback on the Crop Ontology development and use
Lead Institute	● Commits to take overall responsibility for the development and maintenance of one or several specific crop ontologies ● Appoints an ontology curator to execute the commitment

Curator	● Takes executive responsibility on behalf of the Lead Institute(s)
	● Liaises with CO helpdesk to obtain a unique crop identification number, technical guidance (as necessary) and the publication of the crop ontology on the CO website
	● Maintains a wide range of contacts within the crop community and relevant external parties
	● In consultation with the Lead Institute and the (global) Ontology Coordinator forms an expert group to assist with the ontology curation
	● Has overall responsibility for the compilation of the elements of the crop ontology in the appropriate format (Trait Dictionary Template)
	● Responsible to validate mapping to Trait Ontology in collaboration with Technical Coordinator
	● Has overall responsibility for the maintenance of ontology quality standards agreed with the crop community and global standards agreed for implementation across all crop ontologies
	● Provides timely feedback and validation of crop ontology related issues signaled by users or contributors
	● Promotes the use of the crop ontology in the wider user community :
	  ○ training/instruction/public awareness
	  ○ Encourage scientists to use the CO identifiers for traits and variables in the peer-reviewed publications and link to the online crop ontology in the underlying data.

Expert group	● Members make a commitment to assist the curator with the validation and ontology curation in specific areas of expertise ● Promote the use of the crop ontology in the community
Contributors	● Members of the crop community that make specific contributions to the ontology based on their expertise. Their contributions are attributed.
Curator Committee	● Group of Curators representing all crop ontology curators ● Selected by the members of the curator community for a term up to 2 years ● Acts as a standing body to facilitate communication between the entire curator community and the CO Team

The selection of one or more lead institutes is commonly made on the basis of reputation and a (long-)standing involvement with the crop in question. The lead institute(s) take ownership of the development and maintenance of the ontology. This is not an exclusive type of ownership, but a stewardship on behalf of the crop community. This is made explicit by the fact that the intellectual output, the crop ontology, is released under a CC BY 4.0 license which attributes the effort made by the lead institute but allows the broadest possible access and (re-)use by third parties.

The Ontologies are public, released with the Creative Commons license CC- By 4.0.
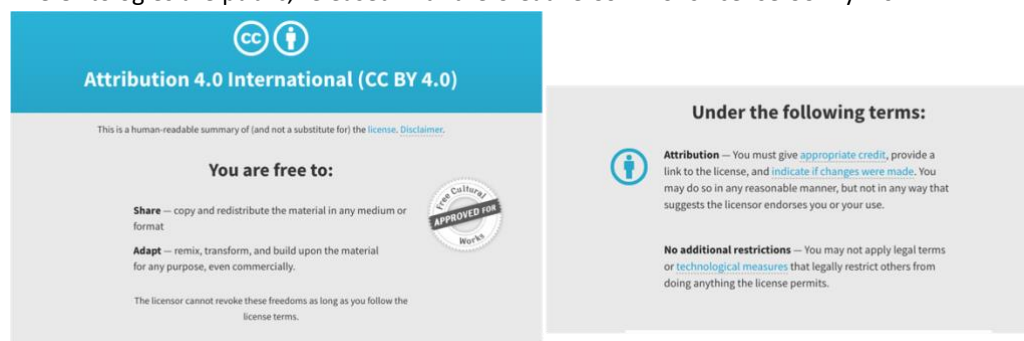
Although lead institute(s) are often thought to be accepting responsibility for an undetermined period, it is good practice to be explicit about the duration. This contributes to overall levels of transparency and sets a milestone for the lead institute(s) to assess their commitment and renew or hand over to an alternative lead when this would be appropriate.

The curator executes the commitment on behalf of the lead institute(s). This executive power should preferably be assigned to a single person. The curator and the members of the expert group should also indicate the duration of their assignment including foreseen options to renew.

The roles of the Lead Institute and the curator are formally acknowledged in the metadata that is shared with Ontology Registries that replicate the crop ontology in question.

ROLES AT THE CROP ONTOLOGY GLOBAL PROJECT DOMAIN

Actors	Responsibilities
CO Management Institute	- Accept responsibility to manage the overall CO project - Appoint the Ontology Coordinator and Technical Coordinator - Provide necessary resources to maintain the web site and develop necessary new features
Ontology Coordinator	- Lead the CO team at global level on behalf on the CO Management Institute - Manage the product, its content, its web site and the quality assurance process - Manage the regular update of the Guidelines and Template, quality checking tools, validation and that curation roles are in place. - Maintain an effective governance framework

	● Regularly consult the community of curators to get feedback and needs ● Interact the Strategy Committee to get timely recommendations on global CO strategy ● Validation of new crops, after checking if the governing roles are in place ● Allocate unique crop codes from CO list ● Provide guidance to users ● Initiate fundraising and engage resources ● Lead the outreach activities with the Ontologies Community of Practice ● Engage community members in decision making and communication
Technical Coordinator	● Member of the CO team ● Update the CO content by providing guidance and expertise to curators ● Promote the use of best practices concerning Crop Ontology development ● Contribute to the discussion around content and CO domain boundaries ● Provide guidance in the use of other ontologies useful in agriculture ● Contribute to the improvement and maintenance of the Template and Guidelines ● Maintain & upgrade, in collaboration with the developer, the CO website backend & frontend according to users' requirements ● Maintain the domain name of the CO website ● Check crop ontologies for quality and inconsistencies ● Operate ontology helpdesk ● Promotion of the CO ● Create the term mappings with Planteome ● Technical interactions with ontology registries, and the breeding databases ● Training of curators
Strategy Committee	● Committee with representatives from crop domains, CO global project domain and external domain (including representatives of Breeding databases and ontology registries) ● Provide recommendations to the CO Ontology Coordinator on strategic project issues
Scientific Advisory Committee	● Committee with representatives from crop domains and external (scientific) domain ● Provide advice to the CO Team (Ontology and Technical Coordinators) on scientific issues relevant to the CO ● Provide guidance and recommendations for the evolution of the Crop Ontology content and technology, identification of best practices regarding the use CO or other ontologies ● Provide expertise and guidance for the homogenization of entities and attributes terms across species – e.g. a tissue name

	- Validate the quality assurance tools proposed by the Community that can be promoted through CO - Validate for some species a group of experts that would be able to evaluate the submissions (homogeneity). The group would make critical assessments (redundancy checking, homogenization of terms...) - Validate the modifications of the Trait Dictionary Template and accompanying Guidelines in consultation with the Ontology and Technical Coordinators

In 2009, Bioversity International took the lead of the CO Project and now the Management Institute is the Alliance Bioversity-CIAT. Although it has taken on this responsibility for an undetermined period, it is desirable the Alliance would state explicitly until when they are committed to act as the CO Management Institute, and for how long the Ontology and Technical Coordinators are appointed. This contributes to overall levels of transparency and sets a milestone to assess its commitments and renew or relinquish them when this would be appropriate. The team coordinating the project provides the helpdesk and can make decisions regarding the update of an ontology when a curator(s) is not yet nominated or when the nominated curator(s) does not provide timely feedback within an agreed deadline.

The Strategy Committee and the Scientific Advisory Committee are called by the CO Global Coordinator. Members are invited after consultation with the Community of Curators and the Technical Coordinator. The invitations will be for initial 2 years that can be renewed as needed.

INTERACTIONS WITH THE EXTERNAL AUDIENCE

The External audience covers all actors which are not in the crop domains or Global Project domain. This includes a wide variety of actors such as, other expert domains, Research Organizations, Donors, End-user groups (Database and Repository Managers, API developers, Database managers) etc.

The external audience is an important source of expertise and insight which is key for aspects ranging from practical implementation to project strategy.

Through the helpdesk and GitHub issues, the external audience can provide useful feedback on the ontology content, format and web site features. Actors of the external audience can subscribe to become members of the Ontologies Community of Practice (https://www.linkedin.com/groups/13707155/) where expert knowledge exchange ontology are regrouped. Representatives of the audience can be invited to attend the Strategy Committee meetings.

LINKAGE BETWEEN THE DOMAINS

With such a diverse set of actors across the three domains, it is of great importance that these domains are well connected to each other.

There are numerous person-to-person contacts between the crop-specific domain(s) and the CO Global Project Domain required to work together on specific ontology-related issues. The Curator Representatives Committee is a standing body to further facilitate the communication between the curator community and CO Team.

In addition, there are regular (virtual) meetings of the Crop Ontology Community of Practice which draws in participants from all three domains (Crop, Project and External Domains). There are also monthly (virtual) meetings of the CGIAR Ontology Working Group which brings together members from across the entire CGIAR ontology community. Both meetings are designed to share information and coordinate activities. It could be worthwhile to consider a mechanism where only the curators and the CO Team meet, at least once a year, to have an opportunity to exclusively discuss internal project matters.

The Strategy and Scientific Advisory Committees also have important roles to play to reach across domains, widen everyone's perspective and promote inter-domain linkages.

QUALITY ASSURANCE PROCESS

Responsibilities for quality assurance have been allocated at the individual crop ontology and global CO Project level (see roles and responsibilities). A crop ontology creation and submission workflow has been designed with quality assessment as an integral part (See Figure 1). Several tools have been developed to assist with the implementation of quality assurance.

CROP ONTOLOGY GUIDELINES

The Guidelines support the use of the Crop Ontology Template to develop the Trait Dictionary (TD) for a new species. Both are updated when changes are required by the Community. Modifications are often suggested by curators – ideally through GitHub issues, and will be validated by the Scientific Advisory Committee. It is strongly recommended reading and applying the Guidelines to develop a high-quality Trait Dictionary. This is the condition to enable the use of the Crop Ontology by a wide community, including industries, and robust mapping with other ontologies. A CO's traits that are properly described following the guidelines are progressively mapped to the Planteome Trait Ontology (TO) maintained by Oregon State University, thus enabling users to search for a trait without consideration of the species for e.g. studies in comparative genomics or for grouping traits for a family or a clade.

TRAIT DICTIONARY TEMPLATE

The TD Template, which is downloadable from the CO website in the 'Guidelines' Menu, is an Excel structured format used to compile, curate, and harmonize the phenotypic variables for the crop and produce a Trait Dictionary that will be published online as a specific ontology. It is composed of a readme tab, a term submission sheet and an embedded Quality Assessment tool.

Any necessary modification of the TD Template is undertaken by the Ontology and Technical Coordinators with the validation by the Scientific Advisory Committee. Modifications will change the version of the Template for clarification. The versioned Template is published in CGSpace along with the corresponding versioned Guidelines (https://cgspace.cgiar.org/handle/10568/110906).

Note: To guide the creation and the update of variables, the Triticeae Toolbox (https://triticeaetoolbox.org/) team, at Cornell University, has developed an intermediate Excel workbook format that holds a worksheet for each ontology data type (one for variable, trait, method, scale). This way repeating terms such as traits, methods or scales have to be entered only once. If you wish to test their format along with the scripts, particularly if you are a user of Breedbase, you can access it on the GitHub: https://github.com/Triticeaetoolbox/ontology-scripts

TRAIT DICTIONARY QUALITY CHECKING TOOLS

A TD quality checking tool is now available in the Trait Dictionary Template. It should be used by both the curators and the Technical Coordinator to check if errors are still present in the Trait Dictionary file. The Quality checking tool returns the list of errors to be fixed but does not impose the modifications.

A PROJECT VERSIONING PLATFORM WITH A REPOSITORY FOR EACH CROP

The use of the Planteome GitHub Platform (https://github.com/Planteome) enables the term submission approval process and the versioning management. A GitHub repository is created in this platform for each crop with an ontology. The curators of the species ontology, the Technical Coordinator, the Planteome Ontology

curators and the development team watch the GitHub repositories and thus receive alert mails about any issues posted by users.

ONLINE SUBMISSION TOOLS FOR NEW TRAITS

When only few traits are requested for a specific crop or some modifications to suggest, contributors can use the online submission forms https://trait-requests.planteome.org/ and https://submit.rtbbase.org/) that will create an issue on GitHub and alert the curators.

Planteome uses protected 'master' GitHub branches for Roots, Tubers and Bananas crops requiring that another curator approve changes to the ontology.

DOCUMENTING QA PROCEDURES

A number of QA tools are in place at two levels: (a) during the Trait Dictionary Development to be used by curators and (b) during the TD upload on the web site by the Technical Coordinator. Direct exchange between the team developing the TD and the Technical Coordinator occurs to validate the suggested changes or request some clarification. IT would be desirable to document the QA process in more detail i.e. which procedures are used, when these are used and how the outcomes are reported and acted upon. Transparency in these matters will further solidify trust in the crop ontologies and add significant value.

BIBLIOGRAPHY

Arnaud E. et al. (2020). The Ontologies Community of Practice: A CGIAR Initiative for Big Data in Agrifood Systems,Patterns, Volume 1, Issue 7, 2020, 100105,ISSN 2666-3899,https://doi.org/10.1016/j.patter.2020.100105 .

Cooper L, Meier A, Laporte M-A, Elser JL, Mungall C, Sinn BT, Cavaliere D, Carbon S, Dunn NA, Smith B, et al (2018) The Planteome database: an integrated resource for reference ontologies, plant genomics and phenomics. Nucleic Acids Research **46**: D1168–D1180

Blake VC, Birkett C, Matthews DE, Hane DL, Bradbury P, Jannink J-L (2016) The Triticeae Toolbox: Combining Phenotype and Genotype Data to Advance Small-Grains Breeding. The Plant Genome. doi: 10.3835/plantgenome2014.12.0099

Jonquet, C., Toulet, A., Arnaud, E., Aubin, S., Dzalé Yeumo, E., Emonet, V., Graybeal, J., Laporte, M.-A., Musen, M.A., Pesce, P., Larmande, P. (2018). AgroPortal: A vocabulary and ontology repository for agronomy, Computers and Electronics in Agriculture, Volume 144, 2018, p. 126-143, doi: 10.1016/j.compag.2017.10.012.

Leonelli, S., Davey, R.P., Arnaud, E., Parry, G., Bastow, R., (2017). Data management and best practice for plant science, Nature Plants, vol. 3, 17086 (2017), DOI: 10.1038/nplants.2017.86

Papoutsoglou, E.A., Faria, D., Arend, D., Arnaud, E., Athanasiadis, I.N., Chaves, I., Coppens, F., Cornut, G., Costa, B.V., Ćwiek-Kupczyńska, et al., (2020). Enabling reusability of plant phenomic datasets with MIAPPE 1.1. New Phytol. doi:10.1111/nph.16544

Selby, P., Abbeloos, R., Backlund, J.E., Basterrechea Salido, M., Bauchet, G., Benites-Alfaro, O.E., Birkett, C., Calaminos, V.C., Carceller, P., Cornut, et al., (2019). BrAPI consortium. BrAPI-an application programming interface for plant breeding applications. Bioinformatics. 2019 Oct 15;35(20):4147-4155. doi: 10.1093/bioinformatics/btz190. PMID: 30903186; PMCID: PMC6792114.

Siverstein K. (2019).'Agricultural Ontologies in Use: The GEMS Platform', BlogPost in CGIAR Platform for Big Data in Agriculture, 2019. https://bigdata.cgiar.org/blog-post/agricultural-ontologies-in-use-the-gems-platform/

Crop Ontology Governance and Stewardship

Virtual Workshop

30 April 2021

Participants

Alliance Bioversity-CIAT	Elizabeth Arnaud
	Alberto Fabio Guerrero
	Marie Angélique Laporte
	Valheria Castiblanco
Bayer Crop Science	Erick Antezana
Boyce Thompson Institute	Naama Menda
CIMMYT	Rosemary Shrestha
	Kate Dreher
CIP	Vilma Hualla
	Elisa Salas
	Thiago Mendes
	Godwill Makunde
CIRAD	David Pot
IBET-ITQB NOVA	Inês Chaves
ICRISAT	Abhishek Rathore
	Roma Das
IITA	Agbona Afolabi
Integrated Breeding Platform	Julian Pietragalla
INRAe	Cyril Pommier
	Celia Michotey
IRRI	Jeffrey Detras
	Kenneth McNally
	Nikki Borja
John Innes Centre	Luzie Wingen
Oregon State University	Laurel Cooper
	Pankaj Jaiswal
Southern Cross University	Liliana Andres Hernandez
	Ramil Mauleon
Washington State University	Jing Yu