Towards a Reference Plant Trait Ontology for Modeling Knowledge of Plant Traits and Phenotypes

Elizabeth Arnaud[1], Laurel Cooper[2], Rosemary Shrestha[3], Naama Menda[6], Rex T. Nelson[5], Luca Matteis[1], Milko Skofic[1], Ruth Bastow[4], Pankaj Jaiswal[2], Lukas Mueller[6] and Graham McLaren[7]

[1]*Bioversity International, via dei Tre Denari, 174/a, Maccarese, Rome, Italy*

[2]*Dept. of Botany and Plant Pathology, Oregon State University, Corvallis, OR 97331, U.S.A.*

[3]*Centro Internacional de Mejoramiento de Maiz y Trigo (CIMMYT), Texcoco, Edo. de México, Mexico*

[4]*GARNet, School of Life Sciences, University of Warwick, Wellesbourne, Warwick, CV35 9EF, U.K.*

[5]*USDA-ARS CICGRU, 1012 Crop Informatics and Genetics Laboratory, Ames, IA 50011, U.S.A.*

[6]*Boyce Thompson Institute for Plant Research, Tower Road, Ithaca, NY 14850, U.S.A.*

[7]*Generation Challenge Program, % CIMMYT, Texcoco, Edo. de México, Mexico*

Keywords: Agriculture, Plant Phenotype, Plant Trait Ontology, Integrated Breeding Knowledge, Community of Practice.

Abstract: Ontology engineering and knowledge modeling for the plant sciences is expected to contribute to the understanding of the basis of plant traits that determine phenotypic expression in a given environment. Several crop- or clade-specific plant trait ontologies have been developed to describe plant traits important for agriculture in order to address major scientific challenges such as food security. We present three successful species and/or clade-specific ontologies which address the needs of crop scientists to quickly access a wide range of trait related data, but their scope limits their interoperability with one another. In this paper, we present our vision of a species-neutral and overarching Reference Plant Trait Ontology which would be the basis for linking the disparate knowledge domains and that will support data integration and data mining across species.

1 INTRODUCTION

Agricultural crop- or clade-specific databases provide comparative phenotypic and genotypic information that helps elucidate functional aspects of plant and agricultural biology. For researchers, it is necessary to have seamless access to various distributed and interrelated data sources such as genetic, trait, genotypic and experimental data to explore biologically interesting questions. Agricultural centers have a huge amount of historical data that reflects a sound scientific knowledge of crop biology and physiology. Plant scientists are producing large volumes of data on genetic mapping, gene expression, and full genome sequences that can be used to gain better insights into plant traits and phenotypes.

Traditionally, phenotype information has been captured in a free text manner, which cannot be easily indexed and presents an obstacle to data sharing. One approach to overcome this obstacle is through the annotation of data using a common controlled vocabulary or “ontology" (Ashburner et al., 2000; Smith et al., 2007). An ontology is a way of representing knowledge in a given domain that includes a set of terms to describe the classes in that domain, as well as the relationships among terms. Each term can be associated with an array of data such as names, definitions, identification numbers, and genes involved. Ontologies are fundamental for unifying diverse terminologies, and are increasingly used by scientists in many fields and by the online web search engines. In an ontology, terms are carefully defined and are related to each other using logically defined relationships as defined by the OBO Foundry Relations Ontology (RO; Smith et al., 2005) and supported by the prevailing knowledge. Such structured ontology trees allow researchers to

Arnaud E., Cooper L., Shrestha R., Menda N., Nelson R., Matteis L., Skofic M., Bastow R., Jaiswal P., Mueller L. and McLaren G..
Towards a Reference Plant Trait Ontology for Modeling Knowledge of Plant Traits and Phenotypes.
DOI: 10.5220/0004138302200225
In *Proceedings of the International Conference on Knowledge Engineering and Ontology Development* (KEOD-2012), pages 220-225
ISBN: 978-989-8565-30-3

use terms consistently in scientific publications or standardized handbooks on quality/trait evaluations, and to search for and integrate data linked to these terms in anatomical, genetic, genomic, and other types of biological databases.

2 THE SEMANTIC LANDSCAPE OF GENOTYPE, PHENOTYPE, AND TRAIT

The concepts of genotype and phenotype are among the most fundamental in all of genetics, developmental and evolutionary biology. Plant breeding, particularly requires the integration of these concepts to understand how and why phenotypic expression varies with the environment. A multidisciplinary approach will help address such complex questions. Crop modeling can play a crucial role and requires knowledge integration, which means that molecular geneticists, physiologists and crop modelers can share their respective 'language' (Wollenweber, 2005) and ontology engineering provides a mean to achieve this.

A **genotype** of an organism is the inherited instructions it carries within its genetic code (i.e. the genome). A genotype can be characterized by sequencing genes, as well as by genetic mapping to characterize variations in the DNA sequence. Not all organisms with the same genotype look or act the same way because appearance and behavior are modified by environmental and developmental conditions. Likewise, not all organisms that look alike necessarily have the same genotype.

A **phenotype** (from Greek *phainein*, 'to show' + *typos*, 'type') is the composite of an organism's observable characteristics such as its morphology, development, biochemical or physiological properties, phenology, behavior, and products of behavior (Wollenweber, 2005). Phenotypes result from the expression of an organism's genes and develop over time as the outcome of cumulative causal interactions between genotype and environment (Malosetti et al, 2011). The phenotype comprises the observable characteristics and the expression of particular traits in a particular organism or organism part. It is a composite of an entity (e.g. fruit) and an attribute (e.g. shape) with a value (e.g. round):

Entity + Attribute = Trait (1)

Entity + (Attribute + Value) = Phenotype (observed) (2)

Example:

fruit + (shape + round) = fruit shape round
-> round fruit

3 ENGINEERING COMMON SEMANTIC FRAMEWORK - A REFERENCE PLANT TRAIT ONTOLOGY

Complex free text descriptions used for phenotypes are almost impossible to index and retrieve in a useful way unless ontological concepts are used for the metadata and text tagging. The semantic problem is that, depending on the plant species, the same trait can be given different names. For instance, the trait term *seed color* is referred to as *kernel color* in maize, *grain* or *caryopsis color* in rice and *pod color* in beans but these are all based on the same phenotypic descriptor- *fruit color*. Data integration and/or mining of plant trait data require the identification of equivalent concepts used by the various agricultural research communities.

Having phenotype data scattered in various online databases using their own vocabularies for annotation prevents the integration and comparison of phenotypic and genetic data between species and even across taxa.

The solution to this problem is the development of a true Reference Plant Trait Ontology (Ref-TO; Figure 1). The basis of the proposed Ref-TO is the existing Plant Trait Ontology (TO) (Figure 2) which will integrate and link many crop- and clade-specific trait ontologies. Initially the focus will be on integrating the three crop or clade-specific ontologies described below, but the long-term goal is to describe traits of all plant species.

The development and expansion of such a universal ontology will rely on the long-term involvement of the various plant research communities for maintaining the species-specific terms and applying the ontological terms to their data annotations.

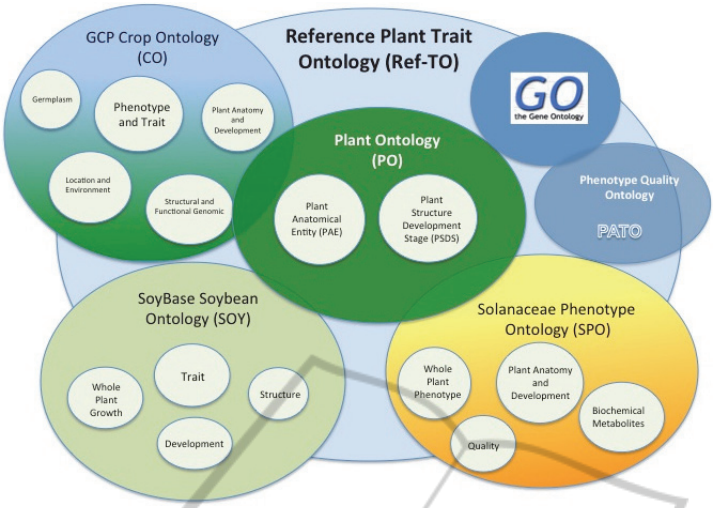

Figure 1: Vision of a Reference Plant Trait Ontology (Ref-TO) to link the crop- and clade-specific trait ontologies.

4 SPECIFIC ONTOLOGIES PROVIDE PLANT PHENOTYPE AND TRAIT INFORMATION

The agricultural research community requires crop trait information, which may contribute to the comparative analysis of genomes and to the selection of promising plant material by crop breeders.

4.1 Soybean Trait Ontology (SOY)

A controlled vocabulary (ontology) describing soybean traits is under active development (SOY), as part of SoyBase (http://soybase.org; Grant et al, 2010), the USDA-ARS soybean genetic and genomic database, a professionally curated biological database for soybean genetic and molecular data. This controlled vocabulary uses terms familiar to the soybean community to facilitate its use. Genetic markers, QTL and soybean gene data are linked to the SOY controlled vocabulary but also cross-referenced to the Plant Trait Ontology (TO) for extension to other crop species. Queries can be initiated at the SoyBase portal using SOY, TO and Plant Ontology (PO) identifiers to access soybean data regarding soybean traits and/or anatomical structures.

In anticipation of further development of the TO, web services are also being developed to allow programmatic access to soybean data using soybean (SOY), Plant Trait (TO), Plant Ontology (PO) or Gene Ontology (GO) identifiers. Semantic web queries of SoyBase data are also available using SSWAP (Gessler et al. 2009) services (Nelson et al. 2010).

4.2 Solanaceae Phenotype Ontology (SPO)

As part of the community-driven SOL Genomics Network (SGN; http://solgenomics.net; Bombarely et al., 2011, Menda 2008), an ontology for Solanaceae phenotypes (SPO) has been developed to describe traits and phenotypes scored by plant breeders in the field. The SGN is a clade-oriented comparative genomic database, focusing on the Euasterid clade, including the Solanaceae family, which has many important crop and model plants such as tomato (*Solanum lycopersicum*), potato (*Solanum tuberosum*), eggplant (*Solanum melongena*), and pepper (*Capsicum annuum*).

Since many Solanaceae phenotype ontology terms are pre-composed, these are also mapped to one or more Plant Ontology (PO) terms, and the Phenotype Quality Ontology (PATO) terms (e.g. the SPO term ‘late fruit ripening’ is mapped to the PATO term ‘delayed’ and PO growth stage term ‘ripening’ and plant anatomy term ‘fruit’).

4.3 The Crop Ontology (CO)

The Crop Ontology (CO) (http://www.cropontology.org/) of the Generation Challenge Programme (GCP) aims to provide a semantic framework to the computational architecture of the knowledge-based system called

the Integrated Breeding Platform (https://www.integratedbreeding.net/).

The CO is designed to provide a structured, controlled vocabulary for the phenotype of important crops for food and agriculture and is collectively developed by various Crop Communities, associated with the centers of the Consultative Group on International Agricultural Research.

The aim is to foster consistency in annotation and to aggregate datasets containing huge amount of historic phenotypic data on a large range of crops not adequately represented in the PO and the TO (Shrestha et al., 2011). Crops currently included are: banana, cassava, common bean, cowpea, groundnut, maize, potato, rice, sorghum, soybean, wheat. Barley, pigeon pea and yam will be added in 2013.

The CO describes agronomic, morphological, physiological, quality, and abiotic and biotic stress related traits of several crops using a number of common relationship types. However, relations were created such as *'method_of', 'scale_of',* and *'derived_from'* to meaningfully describe the traits and their relations to methods and scales.

The CO contributes to the expansion of the Plant Ontology (PO) and to the Plant Trait Ontology (TO), through submission of new terms. Web links between CO terms cross-referenced with major agronomic information sources provide online access to data annotated with similar ontological terms.

The online Crop Ontology is a public resource that acts as an open-source server for names of traits.

5 TOWARDS THE REFERENCE TRAIT ONTOLOGY

5.1 The Trait Ontology (TO) and the Plant Ontology (PO)

International collaborative efforts already exist to develop multi-species ontologies for example the Plant Ontology (PO) and Trait Ontology (TO). The TO (Figure 2) describes phenotypic traits in plants. In its current form, the TO is organized around eight main classes, allowing it to encompass a broad range of plant traits and be species-neutral. The TO is being actively developed in close cooperation with the Plant Ontology, which describes the morphological and anatomical structures of all plants, as well as the stages of development of the plant structures. (Avraham et al, 2008, Jaiswal, 2005, Walls et al, 2012).

TO terms are "precomposed" (Entity-Quality (EQ) form) using terms from the PO and the Gene Ontology (GO), along with other ontologies such as the Phenotype Quality Ontology (PATO), the ontology of Chemical Entities of Biological Interest (ChEBI), the Plant Environmental Conditions Ontology (EO) and the Plant Disease Ontology (PDO; currently under development), as well as others.

The PO itself has been extensively revised over the past two years, with the focus on expanding the scope of the ontology to span all green plants. We can apply the lessons learned from the PO development to developing the TO as a reference trait ontology for all plant species. Further development of the TO is necessary to develop the across-species terms that will be useful to semantically link the crop- and clade- specific ontologies to one another.

5.2 Common Platforms for Data Integration through Web Services

All the crop- and clade-specific ontologies, as well as the PO and TO, are being developed using a common platform, the OBO-Edit software (Day-Richter et al, 2007) developed and promoted by the Gene Ontology (GO; The Gene Ontology Consortium, 2010). This facilitates cross-linking. All these ontologies also use a number of common relationship types. The most common are '*is_a*' and '*part_of*' relations assigned by OBO-foundry (Smith et al, 2005).

The ontologies presented in this paper are available on the BioPortal site of the National Center for Biomedical Ontology (NCBO) (http://bioportal.bioontology.org) for public access, as well as on their respective sites.

Via various processors or extractors, Resource Description Framework (RDF; http://www.w3.org/RDF/) can capture and convey the metadata or information in unstructured (e.g. text), semi-structured (e.g. HTML documents) or structured sources (e.g. standard databases). This makes RDF a perfect solution for representing data that exists in various databases. RDF structures enable synonyms or aliases to be easily mapped to the same types or concepts. This kind of semantic matching is a key capability of the semantic Web. Currently, both the Plant Ontology and Crop Ontology are available in RDF.

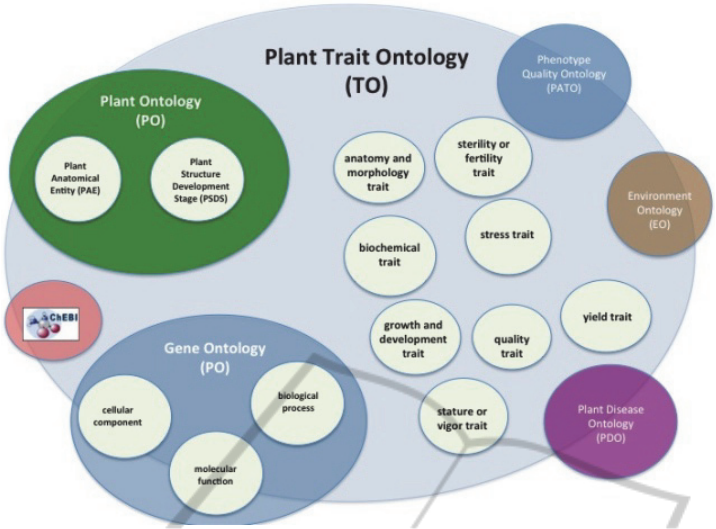

Figure 2: A model of the existing Plant Trait Ontology (TO) showing the species-neutral approach and interaction with the Gene Ontology (GO), the Phenotype Quality Ontology (PATO), the ontology of Chemical Entities of Biological Interest (ChEBI), the Plant Environmental Conditions Ontology (EO) and the Plant Disease Ontology (PDO).

The final objective of a programmatic use of a trait ontology is to support the integration of data sets for given traits, retrieval through web services and the discovery of any piece of information that is annotated with analogous trait concepts. Currently, the existence of many distinct ontologies results in a discontinuous semantic framework. Each ontology is presently taking further steps to use web services to synchronize trait names and OBO files. For example, the GCP crop databases and field books for breeders are synchronized for data annotations with CO through the API.

Developers who wish to use the Plant Ontology in mobile or desktop applications can now access terms, synonyms, definitions, and comments using PO web services. Built with PHP (http://www.php.net/) and modelling aspects of RESTful software architecture (Fielding, 2000), these services provide PO data encoded in JavaScript Object Notation (JSON) format, a widely-used standard for providing data over the internet. The PO plans to continue to develop these web services and envisions the Reference Plant Trait Ontology being offered in a similar way in the future.

6 CONCLUSIONS

A Reference Plant Trait Ontology is necessary to unify the clade- and crop- specific ontologies and provide the semantic framework for querying, reasoning and data mining across the various species databases. Therefore, our objective is to develop the Reference Plant Trait Ontology by improving and expanding the existing Plant Trait Ontology. Our vision for the future development of the Ref-TO is one of an international consortium of the clade- and crop-specific trait ontologies, and would also include representatives from the model plant database groups (such as GARNet, NASC and TAIR for *Arabidopsis*) and representatives from the Plant Ontology and Gene Ontology.

Cross-referencing species-specific terms will unite the ontologies into a network and, by linking plant phenotypes and traits to information, images and documentation across species and even taxa, the community is building a knowledge base with a broad reach, which will be useful to elucidate functional aspects of plant and agricultural biology.

REFERENCES

Avraham, S., Tung, C.-W., Ilic, K., Jaiswal, P., Kellogg, E. A., McCouch, S., Pujar, A., et al. (2008). *The Plant Ontology Database: a community resource for plant structure and developmental stages controlled vocabulary and annotations.* Nucleic Acids Research, 36(suppl_1), D449–454.

Bombarely, A., Menda, N., Tecle, I.Y., Buels, R.M., Strickler, S., Fischer-York, T., Pujar, A., et al. (2011). *The Sol Genomics Network (solgenomics.net): growing tomatoes using Perl. Nucleic Acids Research*, *39*(suppl 1).

Day-Richter, J., Harris, M.A., Haendel, M., The Gene Ontology OBO Edit Working Group, and Lewis, S.

(2007). Obo-Edit – *An ontology editor for biologists*. Bioinformatics 23:2198–2200.

Fielding, Roy T. (2000). *Architectural styles and the design of network-based software architectures* (Book, 2000) [WorldCat.org]. University of California, Irvine.

Gessler, D. D. G., Schiltz, G.S., May, G. D., Avraham, S., Town, C. D., Grant, D., and Nelson R. T. (2009). SSWAP: *A Simple Semantic Web Architecture and Protocol for semantic web services*. BMC Bioinformatics 10:309

Grant, D., Nelson, R. T., Cannon, S.B. and Shoemaker, R.C. (2010). *SoyBase, the USDA-ARS soybean genetics and genomics database*. Nucl. Acids Res. 38 (Suppl.1):D843-D846.

Jaiswal, P., Avraham, S., Ilic, K., Kellogg, E.A., McCouch, S., Pujar, A., Reiser, L., et al. (2005). *Plant Ontology (PO): a controlled vocabulary of plant structures and growth stages. Comparative and Functional Genomics*, *6*(7-8), 388–397.

Jaiswal P, 2011. *Gramene Database: A Hub for Comparative Plant Genomics Plant Reverse Genetics Methods in Molecular Biology*, Volume 678, 247-275,

Jaiswal, P., Ware, D., Ni, J., Chang, K., Zhao, W., Schmidt, S., Pan, X., et al. (2002). *Gramene: development and integration of trait and gene ontologies for rice.* Comparative and Functional Genomics, 3(2), 132–136.

Malesotti M., Ribaut JM, Eeuwijk van E (2011). *The statistical analysis of multienvironment data: modelling genotype-by-environement interaction and its genetic basis. In* Drought phenotyping in crops: from theory to practice, Part I Plant phenotyping methodology. 123-145

Menda, N., Buels, R. M., Tecle, I., & Mueller, L. A. (2008). *A Community-Based Annotation Framework for Linking Solanaceae Genomes with Phenomes*. *Plant Physiology*, *147*(4), 1788–1799.

Nelson, R., Avraham, S., Shoemaker, R., May, G., Ware, D., & Gessler, D. (2010). *Applications and methods utilizing the Simple Semantic Web Architecture and Protocol (SSWAP) for bioinformatics resource discovery and disparate data and service integration. BioData Mining*, *3*(1), 3.

Shrestha, R., Arnaud, E., Mauleon, R., Senger, M., Davenport, G. F., Hancock, D., Morrison, N., et al. (2010). *Multifunctional crop trait ontology for breeders' data: field book, annotation, data discovery and semantic enrichment of the literature.* AoB Plants, 2010.

Shrestha, R., Davenport, G. F., Bruskiewich, R., and Arnaud, E. (2011). *Development of crop ontology for sharing crop phenotypic information*. *In* Drought phenotyping in crops: from theory to practice, eds. P. Monneveux and J.M. Ribaut (Generation Challenge Programme (GCP), c/o CIMMYT, Mexico) 167-176.

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., et al. (2007). *The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration.* Nature Biotechnology, 25(11), 1251–1255.

Smith B, Ceusters W, Klagges B, Kohler J, Kumar A, Lomax J, Mungall CJ, Neuhaus F, Rector A, Rosse C., 2005 *Relations in biomedical ontologies*. Genome Biology;6(5):R46.

Walls, R.L., Athreya, B., Cooper, L., Elser, J., Gandolfo, M.A., Jaiswal, P., Mungall, C.J., et al. (2012). Ontologies as integrative tools for plant science. *American Journal of Botany*, *In press*.

Wollenweber B, Porter JR, Lubberstedt T (2005) *Need for multidisciplinary research towards a second green revolution.* Curr Opin Plant Biol 8: 337-341