Annotating germplasm to Planteome reference ontologies

A semi-automated methodology

Austin Meier, Laurel Cooper, Justin Elser, Pankaj Jaiswal.
Oregon State University
Corvallis, OR. United States
meiera@oregonstate.edu

Marie-Angélique Laporte
Bioversity
Montpellier, France

***Abstract*— An expected use case of plant phenotype ontologies will be the identification of germplasm containing particular traits of interest. If phenotype data from experiments is annotated using ontologies, it makes sense to include annotations to that germplasm source. A lack of standardized data formatting reduces the utility of these data. Standardizing germplasm data, including links to germplasm databases, or distribution locations improves collaboration, and benefits both researchers and the scientific community as a whole. All plant traits contained in the Planteome reference ontologies are searchable, and interconnected through relationships in the ontology. All data annotated to these reference ontologies will be displayed, shareable, and computable through the Planteome website (www.planteome.org) and APIs. This manuscript will discuss the advantages of standardizing germplasm trait annotation, and the semi-automated process developed to achieve such standardization.**

Keywords—ontology; germplasm; data standards; Planteome; automation.

I. Introduction

Original source data comes in many formats, and many different configurations, as different geneticists, breeders, and genebanks use different methods and standards to record germplasm data. Some standards exist, but not every lab, or breeder uses, or is aware of them. In addition to different formatting and storage, individual researchers use different vocabulary to describe their observations. The result of these different data standards is the inability to share or compare data between groups, and possibly more importantly, the inability for computational iterations on said data. While some standards do exist, not everyone adopts them, or has the technological training or expertise to apply them.

The Planteome aims to provide reference ontologies to describe all aspects of plant life, from genotype to phenotype, including environmental stressors. This network of reference ontologies can be used to standardize workflows, and data annotation. The use of standardized vocabulary to describe data allows for interoperability of different data types across different plant species and between different research groups. Since its advent, the use of next generation sequencing platforms has generated significant amounts of data on a wide range of plant species. The use of standards, and controlled vocabularies has enabled computability and access to this data through machines. On the other hand, phenotypic data has become the bottleneck in genetic studies. Without data standards, much of the phenotypic data that is gathered for genetic experiments is used once, and discarded or left to gather dust. By standardizing and annotating this data to a central hub, which is searchable, its usefulness and longevity would increase drastically. A semi-automated method that makes conforming to these standards a less labor intensive task should increase adoption of standards. This lowered barrier will assist those labs/breeders not currently using an existing phenotyping standard, thereby increasing the value of their data.

II. Methods

A. Overview

Annotating germplasm phenotype data using standardized reference ontologies, developed as a semi-automated protocol, is outlined in figure 1. Details of each step are as follows. These methods can be applied either at the source provider's end or by the Planteome curators and other interested projects.

B. Formatting source data

The phenotype and germplasm data is obtained from the original source provider with their permission. The native data format varies by source, and is the place where manual effort must be used to ensure proper data transformation, quality control, and ontology-based annotation. The original data must include three things: a unique identifier for each germplasm entry, name of the evaluated trait, and a phenotype score (observation) for the trait being evaluated. This combination is distinct from an "entity quality" (EQ) statement in that the trait used in the evaluation already contains the entity and quality, and this information contains the actual score of the trait being evaluated. For example: accession number 1124337 (unique identifier) has a leaf color (trait) of "green" (phenotype score). In addition, several other pieces of information are recommended, such as alternative names and synonyms of germplasm, geographic location identifying the place from where the original seed/plant was collected, and the location

where the recorded phenotype was observed. If the trait was observed using an established variable from the species-specific application ontology such as the Crop Ontology Trait Dictionary [3], this information should be contained within the original data as well.

C. *Mapping traits to reference ontologies*

Data points must be given a 'direct annotation' within the Planteome in order to be displayed properly. In the case of germplasm annotations, that direct annotation is to a trait term in the reference Plant Trait Ontology (TO). Breeders, geneticists, and genebanks evaluate different traits in different ways, so one of the first steps in standardizing germplasm annotation is determining which trait is being evaluated by the researcher, and linking the evaluated trait name with the unique trait identifier within the TO (Eg: "pod color" (soybean) = fruit color (TO:0002617). The resulting map file takes the form of a tab-separated table with two columns: one with the trait name used by the individual researcher, a second with the corresponding trait ID number from the TO. This trait map can be created by hand, or by using an automated script (beta). In the case where a measured trait cannot be mapped to the TO because the corresponding trait is not present in the TO, the user is encouraged to request the creation of a new term by creating an "issue tracker" on the Plant Trait Ontology GitHub page (https://github.com/Planteome/plant-trait-ontology/issues). The creation of this issue tracker allows experts from the specific crop to converse with curators of the ontology to resolve the missing term, completing the mapping, and enriching the reference ontology.

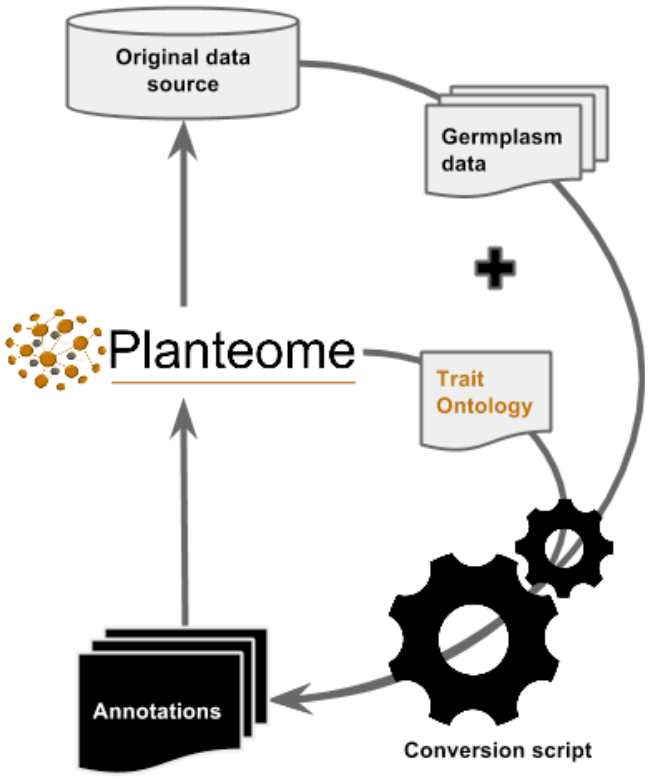

Fig. 1: Workflow outlining the germplasm and phenotype data transformed into standardizing annotations for display on Planteome.org. Original source data, and a trait map mapping observations to terms from Planteome reference ontologies are used by a data transformation program, resulting in standardized annotation files in Gene Annotation Format (GAF2). These files are uploaded to Planteome database, and resulting annotations provide hyperlinks back to the original source.

D. *Conversion script*

The resulting trait map file, combined with the original trait data spreadsheets are used by a Python script (https://github.com/Planteome/common-files-for-ref-ontologies/blob/master/scripts/planteome_germplasm_GAF_translation.py) which converts the data into the standardized GAF2 (Gene Annotation Format 2.0) [1][2] formatted files used to populate the ontology database and display on Planteome.org (Figure 2.) The size and number of these standardized annotation files exceeds the limits imposed by GitHub, and require them to be uploaded to the data repository in SVN (SubVersion). Through SVN, these files are accessible for anonymous bulk download. Users can also choose to download specific annotations directly through the Planteome website.

E. *Cross References to external sources*

In order for a germplasm annotation to link to its original source URL, the host site for the germplasm being annotated must use a predictable URL pattern, in which the unique identifier/accession number is present. Using a combination of the database name, and the unique identifier in the first and second column of the GAF2.0 format annotation files the ontology browser generates a link to the URL in accordance with the pattern specified in the cross references YAML file (https://github.com/Planteome/go-site-xrefs-fork/blob/master/metadata/db-xrefs.yaml).

III. Results

Currently Planteome.org provides germplasm phenotype annotations for nine species: rice, maize, tomato, lentil, cassava, Arabidopsis, eggplant, wild tomato *Solanum chmielewskii,* and *S. pennellii,* totalling 67,763 annotations to ontology terms, over half of which (39,974) have been added through the use of annotation pipelines discussed here. These annotations are linked to traits in the reference Plant Trait Ontology (TO). The Planteome ontology and annotation browser allows complex filtering of these annotations using a number of classifications, and free-text filtering.

IV. Discussion

Plant breeding programs rely on a steady flow of genetic variation into the breeding programs. This increased genetic variation serves no purpose if it does not have associations to useful traits, which can be selected upon. The ability to perform semantic queries of traits of interest to locate germplasm containing these traits enhances the utility of such a platform for plant breeders and geneticists. Planteome aims to become an online informatics portal where breeders can identify traits of interest, and locate data, including

germplasm, QTL, and genes associated with that trait. Having this information centrally located on the web becomes a powerful tool for hypothesis forming, and data sharing, and inter and intra-specific comparisons.

ACKNOWLEDGMENT

This work was supported by IOS:1340112 from the NSF.

REFERENCES

[1] Hill DP, Smith B, McAndrews-Hill MS, Blake J (2008). Gene Ontology annotations: what they mean and where they come from. BMC Bioinformatics 9:S2.

[2] Balakrishnan R, Harris MA, Huntley R, Van Auken K, Cherry JM (2013) A guide to best practices for Gene Ontology (GO) manual annotation. Database. doi: 10.1093/database/bat054

[3] R. Shrestha, E. Arnaud, R. Mauleon, M. Senger, G. F. Davenport, D. Hancock, N. Morrison, R. Bruskiewich, and G. McLaren, "Multifunctional crop trait ontology for breeders' data: field book, annotation, data discovery and semantic enrichment of the literature," *AoB PLANTS*, vol. 2010, p. plq008, 2010.

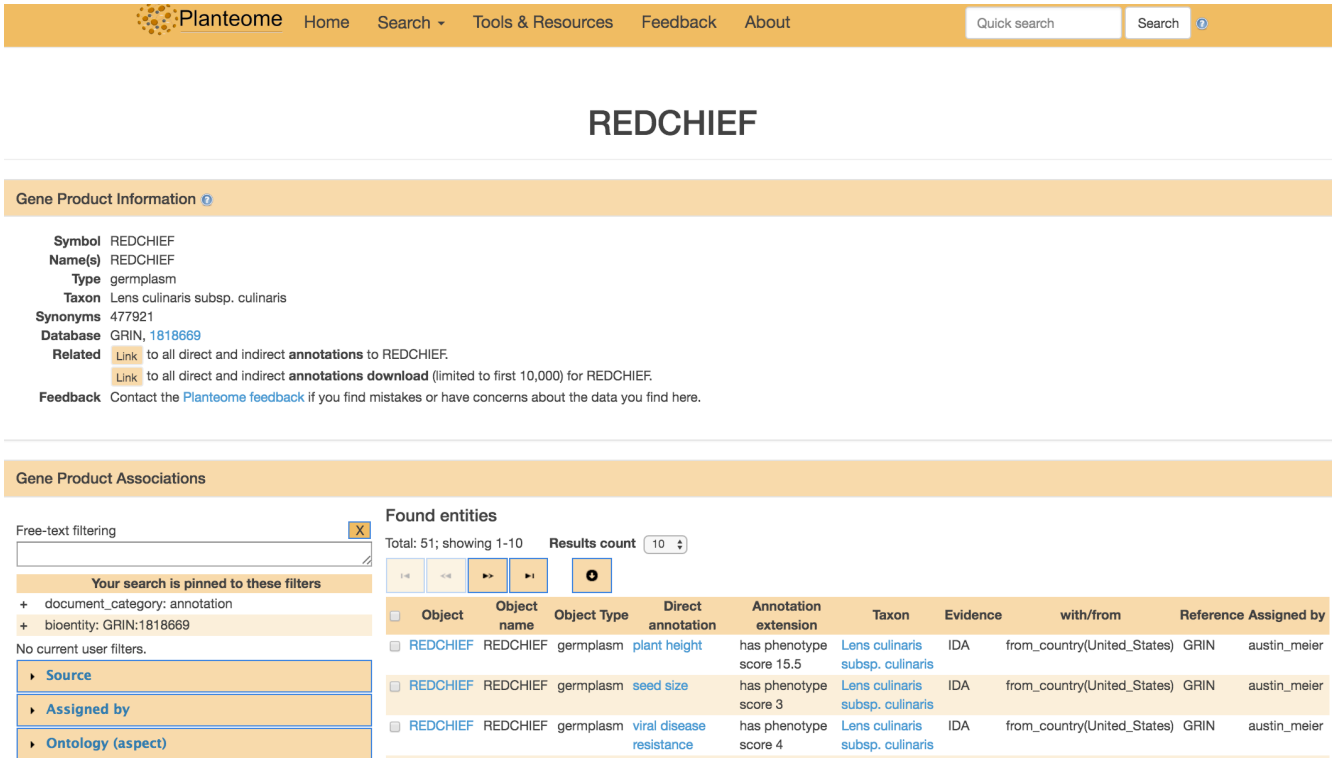

Fig. 2: Screenshot of annotated germplasm displayed on Planteome.org.