OOPS: The Ontology Of Plant Stress

A semi-automated standarization 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

Jorrit H Poelen
400 Perkins Street, Apt. 104
Oakland, CA 94610, USA

***Abstract*— Plant stress traits are important breeding targets for all crop species. Massive amounts of research dollars are spent generating data to combat plant diseases and environmental stress. Often this data is used to achieve a single goal, and then left in a repository to never be used again. As a scientific community, we should be striving to make all publicly funded data reusable, and interoperable. This goal is achievable only through careful annotation using universal data and metadata standards. One such standard is the use of a standardized vocabulary, or ontology. This paper presents a semi-automated method to define and label plant stresses using a combination of web scraping and ontology design patterns. Standardizing the definitions and linking plant stress with established hierarchies leverages previous work of developed knowledge bases such as taxonomic classifications and other ontologies.**

Keywords—ontology; plant pathology; nutrient deficiency; data standards; Planteome; automation; web scraping.

I. Introduction

Global climate change and international travel has introduced more and more diseases to previously unaffected regions. The varieties of crops grown in these regions are typically very susceptible, and yield losses are massive. Spraying pesticides is costly, and damaging to the environment. It takes too long to identify, and integrate resistance genes into existing elite varieties using traditional breeding methods.

Many diseases already have a substantial amount of research and data available related to resistance genes, pathways, and quantitative trait loci (QTLs). However, this data is not easily accessible and even when it is, it can often be difficult to interpret.

By standardizing the naming of plant diseases, their host and pathogen from an ordered taxonomy (e.g. NCBI Taxonomy [1]), and the datasets on genes, QTLs, genetic markers and gene expression, we can ask semantic questions such as: "What genes overlap the resistance QTL, and how they are expressed in response to a pathogen in a given species?", "If the same pathogen affects a closely-related plant hosts, does it trigger the expression of gene homologs?" Or "Is there a common resistance gene motif that is shown to be effective against this pathogen?" Being able to leverage existing datasets will expedite identification of resistance sources, and reduce breeding integration times; producing more food, and using fewer resources. However from the pathology side, using the metadata we can also build a network of ontologies from different knowledge domains to suggest how a stress/disease is manifested. This can be helpful for not just the researchers, but can be integrated into online digital tools to help farmers, agriculture extension specialists, education and machine learning-based data processors for active learning.

II. Methods

A. Overview

The hierarchy of the Ontology Of Plant Stress (OOPS) separates plant stress into two general subclasses: biotic stress, and abiotic stress classes (Fig 1.) The abiotic stress class has two subclasses: plant stress caused by an excess or deficiency of some element. The *biotic stress* class has two children terms, herbivory stress and plant disease. These upper level hierarchy terms are manually curated, and can be adjusted, or added to if the need arises. Initial abiotic stress terms were populated using existing abiotic stress traits found in the Plant Trait Ontology (TO [2]) and initial plant disease terms were identified by scraping the American Phytopathological Society website (www.apsnet.org) using the Samara webscraping application [3]

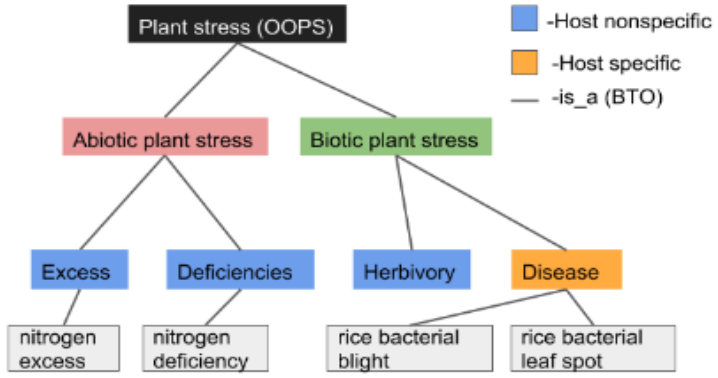

Fig1.
A top level view of the Ontology of Plant Stress (OOPS). All classes fall under the parent class plant stress. The two child terms under the top level divide *plant stress* processes into either biotic stress or abiotic stress. Classes highlighted in blue represent classes in which there is no specificity to the host plant experiencing the stress process. Classes highlighted in yellow indicate stresses in which a specific interaction is occurring between the host plant and the stressor. Example stress classes from table 1 and 2 are displayed in grey.

B. Design patterns

In order to increase automation in development of the Ontology of Plant Stress, we are using a set of design patterns that describe different plant stresses compliant with the Dead Simple OWL Design Patterns (DOS-DPs) format [4]. Using design patterns allows term lists to be maintained in flat tables that can be automatically converted into web ontology language (OWL). In its current pre-release state, OOPS uses three distinct patterns to define plant stress ontology terms: deficiencies, and excess for abiotic stress processes. A single 'disease pattern' is used for biotic stresses.

C. Abiotic stress patterns

Plants can experience stress from exposure to a multitude of different chemical elements, and the process of experiencing stress is dependent on the concentration of said element for a given species or variety of plant in contrast to a reference entity. Abiotic stresses are divided into subclasses based on the excess and deficient states of the stressor element. Stresses caused by exposure to an experimental condition containing too much of an element fall under the "excess" pattern, whereas stresses caused by exposure to an experimental condition that is deficient/lacking a particular element are said to be "deficient". The pattern returns an ontology term with the axioms in Manchester syntax [5] as follows:

Excess pattern:

```
"'abiotic plant stress' and 'causally downstream of' some ('plant treatment' and 'has exposure stimulus' some (ELEMENT and 'has quality' some 'increased amount')) and 'occurs in' some PLANT STRUCTURE"
```

Deficiency pattern:

```
"'abiotic plant stress' and 'causally downstream of' some ('plant treatment' and 'has exposure stimulus' some (ELEMENT and 'has quality' some 'decreased amount')) and 'occurs in' some PLANT STRUCTURE"
```

In the above axioms, the 'ELEMENT' is defined by some entity which is the agent responsible for the stress. This element can be anything, but is typically some chemical entity, defined using Chemical Entities of Biological Interest (ChEBI [6]). The 'PLANT STRUCTURE' is where the stress occurs or is observed, typically defined by a plant anatomy term from the plant ontology (PO [2]), which can be a specific plant part (eg: *root* (PO:0009005), or *vascular leaf* (PO:0009025)), but is often more generally defined as the *whole plant* (PO:0000003). Examples of the tabular list needed to generate both excess stress terms and deficiency stress terms can be seen in Table 1.

Element	Plant Structure
Nitrogen atom (CHEBI: 29352)	whole plant (PO:0000003)
Phosphorus (CHEBI:28659)	whole plant (PO:0000003)
Nitrogen atom (CHEBI: 29352)	leaf (PO:0025034)

Table 1: Flat list describing entities used to construct excess or deficiency plant stress terms in OOPS. Example terms identified from the Plant Trait Ontology terms, nitrogen sensitivity (TO:0000011), and phosphorus sensitivity (TO:0000102). The first column contains the stressor agent, often a chemical entity. The second column contains the anatomical plant structure (from the Plant Ontology) affected by the stress process.

D. Biotic stress patterns

The *Biotic stress* class has two subclasses: herbivory, and plant disease. The Herbivory stress pattern is under development, and the plant disease stress pattern results in the following axiom.

Disease pattern:

```
"'plant disease process' and ('has participant' some HOST) and 'causally downstream of' some ('plant treatment' and 'has exposure stimulus' some PATHOGEN) and 'occurs in' some PLANT STRUCTURE"
```

Defining diseases as processes allows the annotation of stage-specific disease symptoms as infection occurs. Plant diseases are defined by three object classes: host, pathogen, and the plant structure where infection occurs. This pattern defines a host as some participant in the process, whereas the pathogen is said to be an exposure stimulus in an environment containing the pathogen. The disease process is said to occur in some *plant structure* (PO:0009011). This additional requirement allows root diseases to be defined separately from shoot diseases in the case that both are caused by the same pathogen (Table 2). Identification and treatment of diseases depends on the location of the infection. In the cases where the pathogen infection is systemic, *whole plant* (PO:0000003) is used as the plant structure.

Unlike abiotic stresses, plant diseases are processes that are specific to their host plant. It is understood that certain plant pathogens are capable of infecting multiple hosts [9], and this can cause some term inflation within the ontology. This is an acceptable side effect of describing plant stress in as unambiguous terms as possible. Currently, both hosts and pathogens (including pests) are defined by their NCBI taxon ID and are grouped by their taxonomic clade. This allows filtering of diseases based on host, or causal agent (eg: viral diseases vs. bacterial diseases, or potato diseases vs Solanaceae diseases). This will allow potato breeders to filter out all diseases that do not affect potato, or potentially gain insight into resistance mechanisms by expanding the filters to include diseases

affecting all solanaceous crops. Examples of the tabular format needed to generate plant disease terms can be seen in Table 2.

Host	**Pathogen**	**Plant Structure**
Oryza sativa (NCBITaxon:4530)	*Xanthomonas oryzae pv. Oryzicola* (NCBITaxon:1080340)	whole plant (PO:0000003)
Oryza sativa (NCBItaxon:4530)	*Xanthomonas oryzae pv. Oryzicola* (NCBITaxon:1080340)	vascular leaf (PO:0009025)

Table 2: Example rows from the flat list of entities used to generate plant disease terms in OOPS. Three entities are needed: host, pathogen, and plant structure. Both host and pathogen come from NCBI Taxonomy hierarchy, and the plant structure entity affected by the plant disease is from the Plant Ontology.

E. Initial term population

The initial set of abiotic stresses were determined by extracting all of the abiotic plant traits from the Plant Trait Ontology. Any time a plant trait was defined as the response to a chemical entity (ChEBI), two stress terms were created: one each for the excess and deficient state of the said chemical entity.

F. Samara's APS web scrape

To collect plant disease names, the American Phytopathology Society (APS) web publication "Common Names of Plant Diseases" [7], was scraped by the Samara tool [3]. Samara is a command-line tool implement in scala (https://scala-lang.org) that extracts plant trait data from open data sources like APS and USDA-GRIN (www.apsnet.org, www.grin-global.org).

To convert human readable pages from APS's "Common Name of Plant Diseases" resource, an automated process was implemented. The first step of this process is to extract all disease names, source citations, host plant and pathogen from individual host disease pages. The second step corrects troublesome names using a version controlled name map (i.e., nameMap.tsv). The third step links host and pathogen names to NCBI Taxonomy, OBO Relations Ontology (e.g., pathogen of, http://purl.obolibrary.org/obo/RO_0002556) and Plant Ontology for other entities such as host parts (e.g., leaf or root). The relationship, or interaction type, is inferred from the context of the resource and the host parts were extracted from the common name for the disease using a word matching algorithm. The final step exports the results into a tab-separated-value file to make the results available for downstream processing. This process is then repeated to optimize the quality of the name mapping and linking methods.

Given that the APS pages used to extract information were designed for consumption by humans, the structure of the information is not consistent. By providing a rapid, automated process to extract, correct and publish a machine-readable datasets, we put in place a repeatable process in which corrections can be made relatively quickly by avoiding unnecessary manual inputs. For instance, a change in a name mapping file in Samara will automatically trigger a new scrape of the APS resource using a Jenkins job running on a server provided by the Berkeley BBOP [8]. A new dataset will become available less than 20 minutes after that name mapping change is made. Also, dataset archives produced by this automated process are regularly ingested by Global Biotic Interactions (GloBI, https://globalbioticinteractions.org) to further increase the visibility of the APS dataset and the OOPS to stimulate re-use and make it easier to detect suspicious data records.

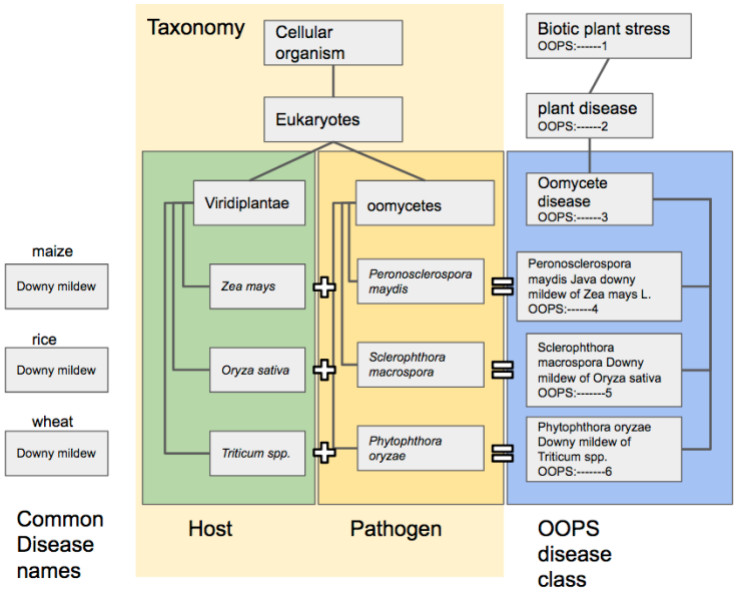

Figure 2: Example differentiation of three plant diseases that previously would be indistinguishable by using only those diseases' common names. By combining the taxonomy of both host and pathogen, we can create unique labels to differentiate between similarly named diseases with completely different causal species.

III. Discussion

The constant arms race between plant hosts, and the pathogens that infect them is guided by evolution - the resulting inference being genes that share similar sequence or domains often share similar functions. OOPS utilizes the relatedness of plant stress participants (host and pathogen in the case of disease, and chemical entity in abiotic stress), and will give scientists improved accuracy when forming hypothesis about gene function, or candidate genes that may be linked to plant traits of interest. Standardizing the definition of plant stresses,

and using this standard vocabulary in the annotation of genes, genomes, QTL, mutants, and the data gathered via field books from plant breeding or field trial experiments can help in building common semantic queries for hypothesis generation, and provide accuracy in the annotation process. Using existing taxonomic hierarchies, and ontologies, researchers can leverage relatedness between both plant hosts, causative pathogens, and even chemical entities to more accurately predict targets for molecular markers, and identify candidate stress responsive gene functions. These standards will also help aggregate existing data, and assist in future-proofing new data to ensure that the massive amounts of both phenotypic and genotypic data being generated can be interoperable instead of being used for an singular task, and dumped into a repository to collect dust.

The real innovation and advancement of this work is the emphasis on automation. Much of the accuracy of the disease terms require information from a subject matter expert. These experts are often not familiar with ontologies and various formats like OWL and ontology editing tools, and would require extensive training and guidance in order to contribute. Therefore, the use of design patterns to automate ontology development, term addition, and edits, allows curators, and contributors to maintain OOPS using just a flat list. This lowered bar for ontology curation reduces effort in training new contributors, additional curators, and the overall overhead for maintenance. Efforts to simplify the construction and maintenance will also improve community involvement and adoption.

Construction of an ontology requires expert domain knowledge to ensure accuracy of the resulting hierarchy. OOPS is no exception. Plant stress spans the entirety of the plant science field, and a single person cannot hope to understand and capture all of the instances of plant stress. That is part of the benefits of using these automated tools for developing an ontology; when issues arise, or additional parental classes are needed to further group stress, they can simply be added to the upper level hierarchy list, and the reasoner can place child terms using the appropriate pattern.

As it currently stands, OOPS is available on GitHub (https://github.com/Planteome/ontology-of-plant-stress). However, it is under construction, and no stable release is available at this time.

IV. Future direction

Community involvement is key to ontology utility. To make OOPS more robust and functional, we are planning to implement a table editing tool that will be accessible to the public. Some form of version control (likely GitHub) will be used to produce robust versioning of stress term edits. Reaching out to subject matter experts, such as CGIAR Research Centers will be key to accurate plant disease descriptions. Reaching out to APS will be important for widespread adoption, and community efforts needed to stay up to date on plant disease nomenclature, and identification. For instance, we imagine a collaboration in which APS updates the Common Names of Plant Diseases [7] pages such that taxonomic terms (host, pathogen) and diseases are linked to NCBI Taxonomy and OOPS respectively, and make them available in formats that are friendly to humans (e.g., html) and machines (e.g., tsv, rdf). In addition, after the release of a stable OOPS, the intent is to link it to the Plant Trait Ontology by using OOPS terms within TO stress responsivity traits. This way, TO, PO, NCBITaxonomy, and ChEBI can all be linked together, to form a more robust knowledge graph within Planteome.

Acknowledgment

This work was supported by IOS:1340112 from the National Science Foundation.

References

[1] Federhen S. The NCBI Taxonomy database. *Nucleic Acids Research*. 2012;40(Database issue):D136-D143. doi:10.1093/nar/gkr1178.

[2] Cooper L, Meier A, Laporte M-A, Elser JL, Mungall C, Sinn BT, Cavaliere D, Dunn NA, Smith B, Qu B et al.. 2018. The Planteome database: an integrated resource for reference ontologies, plant genomics and phenomics. *Nucleic Acids Research.* 10.1093/nar/gkx1152. Vol 46:D1168-1180

[3] Jorrit Poelen, & Marie-Angélique Laporte. (2018, May 7). jhpoelen/samara v0.2.0 (Version v0.2.0). Zenodo. http://doi.org/10.5281/zenodo.1243234).

[4] Osumi-Sutherland D, Courtot M, Balhoff J.P., Christopher Mungall C. Dead simple OWL design patterns. Journal of Biomedical Semantics 2017 8:18. https://doi.org/10.1186/s13326-017-0126-0

[5] Hitzler P, Krötzsch M, Parsia B, Patel-Schneider PF, Rudolph S, (eds).OWL2 Web Ontology Language: Primer: W3C Recommendation; 2009. Available at http://www.w3.org/TR/owl2-primer/.

[6] Hastings J, Owen G, Dekker A, Ennis M, Kale N, Muthukrishnan V, Turner S, Swainston N, Mendes P, Steinbeck C. (2016). ChEBI in 2016: Improved services and an expanding collection of metabolites. *Nucleic Acids Res.*

[7] Common names of plant diseases : American Phythopathology Society. http://www.apsnet.org/publications/commonnames/Pages/default.aspx

[8] http://build.berkeleybop.org/view/Planteome/job/extract-apsnet-diseases

[9] Gilbert and Webb, Phylogenetic signal in plant pathogen–host range *PNAS* 2007. 104 (12) 4979-4983

[10] Cooper and Jaiswal, The Plant Ontolgy: A Tool for Plant Genomics. Methods in Molecular Biology. Vol 1373