AG2PI SEED GRANT PROPOSAL

Title of Proposal:

Plant Stress Ontology: Data Standards and Knowledge Graph

Lead PI (Name, Title, Affiliation(s), email)

Pankaj Jaiswal, Professor, Oregon State University, PI, Planteome reference Ontology project, jaiswalp@oregonstate.edu

Co-PI (Name, Title, Affiliation(s), email)

Graham King, Executive Director, DivSeek International Network; Professor, Southern Cross University, Australia, graham.king@divseekintl.org

Collaborators (Name, Title, Affiliation, email):

Elizabeth Arnaud, Crop Ontology Lead, Data Scientist, Alliance of Bioversity, CIAT, CGIAR, E.ARNAUD@cgiar.org

Chris Mungall, Co-PI Gene Ontology project, Department Head, Biosystems Data Science, Lawrence Berkeley National Laboratory, CJMungall@lbl.gov

Barry Smith, Professor, Director, Ontologist, National Center for Ontological Research, University of Buffalo, ifomis@gmail.com.

Laurel Cooper, Planteome Project Coordinator, Research Associate, Oregon State University, cooperl@oregonstate.edu

Fiona McCarthy, Professor & Geneticist, Animal and Comparative Biomedical Sciences, University of Arizona. Investigator and Data Scientist AgBase. fionamcc@arizona.edu

Grant Administrator:

Office for Sponsored Research and Award Administration

Phone: 541-737-5708

sponsored.pgrograms@oregonstate.edu

Keywords:

ontology, data standards, plant stress, plant disease, knowledge graph

Project Description:

Objectives/aims: We propose to develop a Plant Stress Ontology (PSO) to standardize a controlled structured vocabulary that is used to describe plant stress responses and adaptations. Plant stress refers to any environmental or biological factor that impairs the normal functioning and growth of a plant. These stressors can include abiotic factors such as extreme temperatures, drought, and soil salinity, as well as biotic factors such as pests, pathogens, mutualists and diseases. The typical hierarchical lists of plant diseases and stress, for example those provided by the plant disease bulletins, agriculture extension programs, and plant disease compendiums of the American Phytopathological Society (APS) [1, 2] are a rich source of information. However, they are difficult to cite, explore, and use for automation, exacerbating tracking of changes in the context of evolving information. Therefore a standardized PSO and its edited version tracking will provide the basic framework for developing a common language for describing plant stresses. The species agnostic reference PSO will provide a standardized way to describe and classify different types of plant stresses, their manifestation, measurements, observations, affected plant parts and growth stages, and curated images, phenotype data tables and known molecular interactions, as conceptualized by Walls et al. [3]. This will allow accurate description and comprehensive data collection and analysis. The PSO development and adoption will provide consistency in annotation and data collection in phenomics projects, enable open discussion on sharing information about plant stress responses observed/recorded by different research groups and improve interoperability among online databases. PSO knowledge graph (KG) will also help in unifying similar concepts and diverse vocabularies used in projects on agriculture extension [4], plant ecology, plant genetics, and plant breeding, as well as for training machine learning tools for stress detection [5-9]. We will organize two week-long hands-on workshops by inviting plant pathologists, physiologists and ontology designers to help guide the discussion, build strategies and guidelines leading to development of first versions of the PSO ontology for biocuration of genes, QTLs and genome to phenome training data. Each workshop will fund participation of 10 experts through nomination and targeted crop representation. Additional open invitations will be available to experts. The project aims are:

Aim 1: Develop a Plant Stress Ontology (PSO) framework based on the OBO Foundry principles [10, 11]. PSO will consist of two major classes on abiotic and biotic stresses and will interlink additional branches with causal agents of stress (pathogens, race, drought, salinity, etc.)

to the host plant species/genotype, the plant diseases, mimicry, any common and unique symptoms/observations [1], known methods for scoring phenotypes [12-14] and suggested species/crop-specific growth environment conditions in which a stress response may occur. In the proposed PSO pilot, the workshop participants will describe the five most important biotic and abiotic stresses and corresponding responses affecting the important food crops: rice, maize, soybean, potato and cassava. Compared to simple hierarchical lists of stress names [2], the PSO will be enriched by re-using terms from other ontologies and ontological relationships connecting the host (crop plant) and for biotic stresses, evolving pathogen and vector, disease nomenclature, spatial and temporal aspects of infection/start of stress affecting plant parts and growth stages and most importantly the different levels and types of observed traits in the form of symptoms and responses for example in Figure-1 [3]. The adoption of unique term IDs for each defined stress ontology term and their one-to-many term-to-term relations will help connect phenotypic symptoms that are unique and common to different stresses. For example, a leaf yellowing symptom can be connected to Nitrogen deficiency and rice sheath blight disease. The PSO will also adopt and encourage standardized annotation of host germplasm/accessions and pathogen genotypes/races, QTL and gene-gene interaction, etc. We anticipate supplementing human knowledge with some automation (Natural Language Processing) to process currently available stress descriptors and observations recorded in free text. We will also engage our major stakeholders from the CGIAR Crop Ontology developers [15-17], DivSeek partners, agriculture extension faculty, and the APS members in the PSO development.

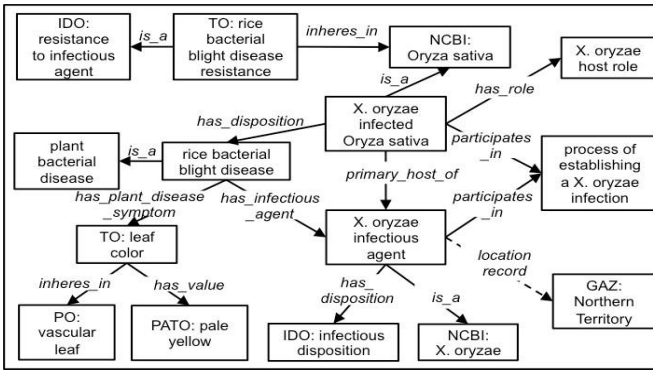

Figure-1: The text descriptions like "disease: rice bacterial leaf blight disease | host species: *Oryza sativa* (rice)| caused by: *Xanthomonas oryzae* | has symptom: pale yellow leaves | reported in: Northern Territory of Australia" can be converted into a conceptualized PSO graph. (Source: Walls et al 2012) [3]. Various sister ontologies like NCBI taxonomy, Trait Ontology (TO), Plant Ontology (PO), Phenotype Attribute Ontology (PATO) will be used to process text descriptions into a PSO KG enabled by multiple parent-child relations.

Aim 2: Develop a knowledge graph (KG) database and tools: We will integrate the PSO into the Planteome Ontology database [18], making improvements to the existing ontology browser to navigate the KG. The PSO and the KG with curated annotations will also be available

in the standardized XML and other formats defined by the OBO foundry [10]. Data APIs will be provided for use in remote applications. A small demonstration set of genes, QTL, phenotype and mutant data will be curated to test the abilities of the KG to answer biological questions, common and unique symptoms and genetic components involved in abiotic and abiotic stress response. Another measure of success will be to see if the generic data model is capable of enriching new stress types and phenotyping methods adopted for studying plant species beyond those listed in this proposal and for adoption by non-plant Ag2PI community.

Aim 3: Develop guidelines and Standard Operating Procedures for enriching PSO for a wider range of species and stresses, which will help record and find species/genotype differences and their varying degree of responses to similar stresses, including common pathogens. We will develop training protocols and guidelines for ontology-based phenotyping of plant diseases and disorders caused by stresses such as drought, heat, flooding or salinity.

2. Furthering the aims of the AG2PI: The goals of the AG2PI project are to addresses agricultural challenges, from genome to phenome, by building and supporting a cross-kingdom and multi-disciplinary research community. This proposal addresses those goals by providing a framework and tools for the collection and utilization of the extensive data generated by the high-throughput phenotyping, genotyping and germplasm evaluation studies, as well as the accumulated legacy data that is available. There is a need to capture, archive, annotate and mine this data in a standardized manner using ontologies, and enable interoperability of online resources for comparative query and analysis. Such resources will allow researchers to query and develop strategies for building improved stress resilient crops through accelerated breeding programs and genetic engineering. The presentations at the annual Plant and Animal Genome, NAPPN and APS conferences will invite users to demo sessions to test the knowledge graph. The Breeding Management Systems will be invited to test PSO integration in their field books and decision making tools. The parts of the KG can be adopted to develop tools for decision making and aid early detection, reporting and mitigation by farmers, and agriculture extension programs at state, national and international-levels. The proposed species/clade agnostic data structure for PSO will be developed for future adoption by other communities, and rigorously tested by the collaborator AgBase [19, 20], a livestock genome-to-phenome database.

3. Expected outcomes & deliverables: We will organize two hands-on workshops to develop PSO, knowledge graph, annotations, guidelines, SOPs, and to curate demonstration public datasets. We will collect training data from (1) the Jaiswal Lab on abiotic stress physiology at Oregon State Univ. and (2) collaborate with NAPPN and CGIAR projects to obtain additional abiotic and biotic stress phenotype measurement methods and observations including those on crop diseases. Ontology and curated datasets will be shared with the Ag2PI community for (1) training and education, (2) developing ML-based decision making tools [5-9], and (3) test the reference PSO in its native form for integration in the OMICs databases by standardizing the plant stress description, nomenclature, capturing methodologies for phenotyping [12, 13]. Livestock databases like AgBase [20] and others from the AgBioData consortium [21] will be invited to test the framework to develop a comparable ontology for their livestock research community. All the data and the documentation will be released to the public from the GitHub and its available disseminated amongst AG2P, DivSeek and other communities. The large annotated data will be served from the project's SVN repository. The ontology database and browser will be hosted by the Planteome [18]. The data APIs will be available for integration in third party tools for remote access. The PIs and collaborators will make project presentations and write peer-review publications, and draw a plan to write a future international collaboration proposal for submission to the NSF/USDA-NIFA Ag2PI and other national/international programs.

4. Qualifications of the project team: Each project team member will be directly involved in one or more activities of ontology design, best practices, training, database development, writing papers and helping recruit participants, besides promoting the project. Pankaj Jaiswal is the project lead of the Planteome project [18] on common reference ontologies for plant biology. He co-founded the Plant Ontology Consortium [22, 23] that led the development of reference ontologies for plant anatomy, plant growth and development, and phenotypic traits. These reference ontologies are now global standards for integration and annotation of the plant genomics data for gene function, expression and phenotype. Elizabeth Arnaud is the Crop Ontology project [16] lead for CGIAR. She leads the development and maintenance of Crop Ontology for 36 major global crops which are used to standardize phenotype data collection via electronic field books and breeding management systems. Graham King is a crop geneticist and Executive Director of DivSeek International Network, and led development of the Compositional

Dietary Nutrition Ontology [24]. His leadership at DivSeek and Southern Cross University, Australia, is playing a major role in encouraging adoption of global data standards in genome-to-phenome projects and data archives. Chris Mungall is the Bioinformatics and Data Science Lead for the popular Gene Ontology Consortium [25] and Open Biomedical Ontology Foundry. He leads the Ontology design, modeling, and developing standards and databases used globally and across the Biological Ontology domain area. Barry Smith, a philosopher and world-renowned ontologist, introduced the concept of Biological Ontologies to the world that led to formation of OBO Foundry. His expertise is ontology design, grammar, and modeling [26]. Laurel Cooper has led the ontology development and training, and coordinated the Plant Ontology and Planteome projects [18] for the past 14 years. Fiona McCarthy leads the AgBase bioinformatics and genome database for livestock [19, 20]. She is familiar with using and developing ontologies for agricultural species and integrated them in livestock genome-to-phenome annotation databases.

5. Proposal timeline

Activity	Q1	Q2	Q3	Q4
Monthly group meetings				
Infrastructure development and maintenance				
Stress Ontology guidelines and framework	Alpha v1	Beta v1	Beta v2	Public V1
Workshop-1 (biotic stress)		Biotic		
Workshop-2 (abiotic stress)			Abiotic	
Presentation at annual PAG, APS & NAPPN Conferences				
Ontology and Annotation Data release		Alpha	Beta	Public V1
Ontology database		Dev v1	Dev v2	Public V1
Publications				
Future proposals				

6. Engaging AG2P scientific communities & underrepresented groups: The AgBioData resources hosting genomics and genetic data, including the national and international gene banks will be encouraged to test the PSO in their data annotation framework. Similarly, we will engage the Ag2PI community in two ways, (1) by using the PSO knowledge graph on the project site, show its benefits in answering biological questions and datamining, and (2) encourage researchers to integrate PSO in the genome-to-phenome data and any software tools developed by their projects. We will encourage and support the participation of underrepresented and minority groups by inviting them to the two project workshops.

Bibliography/References cited

1. MB, R., W. MR, and M. O. *Plant Disease Diagnosis*. 2002 [cited 2022; Available from: https://www.apsnet.org/edcenter/disimpactmngmnt/casestudies/Pages/PlantDiseaseDiagnosis.aspx.
2. APS. *Common names of Plant Diseases*. 2022; Available from: https://www.apsnet.org/edcenter/resources/commonnames/Pages/default.aspx.
3. Walls, R., B. Smith, E. Justin, G. Albert, D. Stevenson, W., and P. Jaiswal, *A plant disease extension of the Infectious Disease Ontology*, in *Proceedings of the Third International Conference on Biomedical Ontology (CEUR 897)*. 2012. p. 1-5.
4. Florence, J. and J.W. Pscheidt, *Mummy Berry Management in the Pacific Northwest.* 2015. **EM 9117**. https://catalog.extension.oregonstate.edu/sites/catalog/files/project/pdf/em9117.pdf
5. Mutka, A. and R. Bart, *Image-based phenotyping of plant disease symptoms.* Frontiers in plant science, 2014. **5**: p. 734.
6. Niederkofler, A., S. Baric, G. Guizzardi, G. Sottocornola, and M. Zanker, *Knowledge Models for Diagnosing Postharvest Diseases of Apples*. 2019.
7. Vo, A., N. Nguyen, T. Nguyen, T. Dang, D. Nguyen, and B. Thien, *An Automatic Recommendation System for Plant Disease Treatment*. 2022. p. 625-637.
8. Hewarathna, A., V. Palanisamy, J. Charles, and S. Thuseethan, *Deep Hybrid Learning Framework for Plant Disease Recognition*. 2022. 49-54.
9. Vats, S. and V. Chivukula, *Plant Disease Detection Using DeepNets and Ensemble Technique*. 2022. 1-6.
10. Jackson, R., N. Matentzoglu, J.A. Overton, R. Vita, J.P. Balhoff, P.L. Buttigieg, S. Carbon, M. Courtot, A.D. Diehl, D.M. Dooley, W.D. Duncan, N.L. Harris, M.A. Haendel, S.E. Lewis, D.A. Natale, D. Osumi-Sutherland, A. Ruttenberg, L.M. Schriml, B. Smith, C.J. Stoeckert, Jr., N.A. Vasilevsky, R.L. Walls, J. Zheng, C.J. Mungall, and B. Peters, *OBO Foundry in 2021: operationalizing open data principles to evaluate ontologies.* Database (Oxford), 2021. **2021**. https://www.ncbi.nlm.nih.gov/pubmed/34697637
11. Smith, B., M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L.J. Goldberg, K. Eilbeck, A. Ireland, C.J. Mungall, O.B.I. Consortium, N. Leontis, P. Rocca-Serra, A. Ruttenberg, S.A. Sansone, R.H. Scheuermann, N. Shah, P.L. Whetzel, and S. Lewis, *The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration.* Nat Biotechnol, 2007. **25**(11): p. 1251-5. https://www.ncbi.nlm.nih.gov/pubmed/17989687
12. Phenospex. *Plant Parameters (Phenna 1.0 and 2.0)*. 2022 [cited 2022; Available from: https://phenospex.helpdocs.com/plant-parameters.
13. Photosynq. *Photosynq References and Parameters*. 2022; Available from: https://help.photosynq.com/view-and-analyze-data/references-and-parameters.html.
14. TerraRef. *Terraref user manual*. 2021; Available from: https://github.com/terraref/documentation/tree/master/user-manual/data-products.
15. Pan, Q., J. Wei, F. Guo, S. Huang, Y. Gong, H. Liu, J. Liu, and L. Li, *Trait ontology analysis based on association mapping studies bridges the gap between crop genomics and Phenomics.* BMC Genomics, 2019. **20**(1): p. 443. https://www.ncbi.nlm.nih.gov/pubmed/31159731

16. Shrestha, R., L. Matteis, M. Skofic, A. Portugal, G. McLaren, G. Hyman, and E. Arnaud, *Bridging the phenotypic and genetic data useful for integrated breeding through a data annotation using the Crop Ontology developed by the crop communities of practice.* Front Physiol, 2012. **3**: p. 326.https://www.ncbi.nlm.nih.gov/pubmed/22934074
17. Shrestha, R., 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, 2010. **2010**: p. plq008. https://www.ncbi.nlm.nih.gov/pubmed/22476066
18. Cooper, L., A. Meier, M.A. Laporte, J.L. Elser, C. Mungall, B.T. Sinn, D. Cavaliere, S. Carbon, N.A. Dunn, B. Smith, B. Qu, J. Preece, E. Zhang, S. Todorovic, G. Gkoutos, J.H. Doonan, D.W. Stevenson, E. Arnaud, and P. Jaiswal, *The Planteome database: an integrated resource for reference ontologies, plant genomics and phenomics.* Nucleic Acids Res, 2018. **46**(D1): p. D1168-D1180. https://www.ncbi.nlm.nih.gov/pubmed/29186578
19. McCarthy, F.M., S.M. Bridges, N. Wang, G.B. Magee, W.P. Williams, D.S. Luthe, and S.C. Burgess, *AgBase: a unified resource for functional analysis in agriculture.* Nucleic Acids Res, 2007. **35**(Database issue): p. D599-603.
20. McCarthy, F. *AgBase 2.0*. 2021; Available from: https://data.nal.usda.gov/dataset/agbase.
21. Harper, L., J. Campbell, E.K.S. Cannon, S. Jung, M. Poelchau, R. Walls, C. Andorf, E. Arnaud, T.Z. Berardini, C. Birkett, S. Cannon, J. Carson, B. Condon, L. Cooper, N. Dunn, C.G. Elsik, A. Farmer, S.P. Ficklin, D. Grant, E. Grau, N. Herndon, Z.L. Hu, J. Humann, P. Jaiswal, C. Jonquet, M.A. Laporte, P. Larmande, G. Lazo, F. McCarthy, N. Menda, C.J. Mungall, M.C. Munoz-Torres, S. Naithani, R. Nelson, D. Nesdill, C. Park, J. Reecy, L. Reiser, L.A. Sanderson, T.Z. Sen, M. Staton, S. Subramaniam, M.K. Tello-Ruiz, V. Unda, D. Unni, L. Wang, D. Ware, J. Wegrzyn, J. Williams, M. Woodhouse, J. Yu, and D. Main, *AgBioData consortium recommendations for sustainable genomics and genetics databases for agriculture.* Database (Oxford), 2018. **2018**.
22. Walls, R.L., L. Cooper, J. Elser, M.A. Gandolfo, C.J. Mungall, B. Smith, D.W. Stevenson, and P. Jaiswal, *The Plant Ontology Facilitates Comparisons of Plant Development Stages Across Species.* Front Plant Sci, 2019. **10**: p. 631. https://www.ncbi.nlm.nih.gov/pubmed/31214208
23. Cooper, L. and P. Jaiswal, *The Plant Ontology: A Tool for Plant Genomics.* Methods Mol Biol, 2016. **1374**: p. 89-114.https://www.ncbi.nlm.nih.gov/pubmed/26519402
24. Andrés-Hernández, L., K. Blumberg, R.L. Walls, D. Dooley, R. Mauleon, M. Lange, M. Weber, L. Chan, A. Malik, A. Møller, J. Ireland, L. Segovia, X. Zhang, B. Burton-Freeman, P. Magelli, A. Schriever, S.M. Forester, L. Liu, and G.J. King, *Establishing a Common Nutritional Vocabulary - From Food Production to Diet.* Frontiers in Nutrition, 2022. **9**. https://www.frontiersin.org/articles/10.3389/fnut.2022.928837
25. Consortium, T.G.O., *The Gene Ontology resource: enriching a GOld mine.* Nucleic Acids Research, 2020. **49**(D1): p. D325-D334.https://doi.org/10.1093/nar/gkaa1113
26. PhilPeople. *Barry Smith*. 2022; Available from: https://philpeople.org/profiles/barry-smith.