Semantic Web 0 (0) 1
IOS Press

Bringing phytopathology onto the reasoned Semantic Web: the Plant-Pathogen Interactions Ontology (PPIO)

Alejandro Rodríguez Iglesias [a,*], Mikel Egaña Aranguren [a], Alejandro Rodríguez González [a] and Mark D. Wilkinson [a]

[a] *Biological Informatics Group, Centre for Plant Biotechnology and Genomics (CBGP), Technical University of Madrid (UPM), Spain*

Abstract. Interactions between plants and plant-pathogenic bacteria have both scientific and economic importance, and are particularly relevant in the domain of biodiversity. While semantically-oriented resources exist that describe certain aspects of plant phenotypes, biodiversity, and plant disease, few are designed specifically to be used together with rich semantic reasoning, and none are specifically aimed at describing the interplay between plants and the organisms that infect them. We present here the Plant-Pathogen Interactions Ontology (PPIO), whose axiomatic models allow the integration of, and inference over, plant-pathogen interaction datasets in a semi- or fully-automated manner.

Keywords: plant pathogenic bacteria, ontologies, Semantic Web, PPIO, plant pathology, reasoning

1. Introduction

Plants can be susceptible to attack by a wide range of pathogenic bacterial genera [12]. If the plant's defense barriers are overcome, the infection process(es) can ultimately result in the death of the plant. Even if the plant survives, pathogenic processes may have significant effects on economically important traits such as crop yield, and as a consequence there is significant interest in capturing and curating plant/pathogen interaction data [13]. Unfortunately, plant-pathogen data is largely collected in a highly distributed manner, by many research teams worldwide, and sometimes even by individual land-workers themselves. As a result, there is a large amount of both structured and unstructured, curated and non-curated, data in this domain of interest, as exemplified by the hundreds of publications that focus their interest on unveiling the mechanisms of pathogenic bacteria interactions with their hosts [6] [7]. Therefore, construction and integration of unified data sources related to plant interactions with pathogenic bacteria is an essential next-step, and will ultimately lead to a better preservation of worldwide crops.

The field of plant-pathogenic bacteria also provides a stark example of the increasing relevance of biodiversity research in the area of plant pathology [3]. Traditionally, *Agrobacterium, Erwinia, Pseudomonas and Xanthomonas* were considered the four primary plant pathogenic genera. With the rapid improvement of research experimental approaches, the number of identified plant pathogenic bacteria genera has increased to at least 30. As a result, large amount of largely unexplored biological data have been created, and these are archived over a wide range of both structured and unstructured resources. It is, clearly, highly desirable to pursue initiatives that will make these data easier to integrate, explore, and interpret.

The effectiveness of Semantic technologies to manage large, distributed datasets has been demonstrated in other areas of the life sciences. However, to date, Semantic Web tools have not been extensively applied

*Corresponding author. Email:alejandroriglesias@gmail.com

to the knowledge domain of plant-pathogenic bacteria. Here we present the Plant Pathogen Interactions Ontology (PPIO), an ontology developed to integrate and organize data about interactions between plants and their (currently bacterial) pathogens. The main goal of PPIO is to serve as a reference for expert plant pathologists by providing the knowledge necessary to assist in their i nterpretation of, and prediction of, the phenotypic responses that result from pathogenic biological interactions.

2. Modelling

The initial data upon which the PPIO was modelled was collected manually by consulting and "scraping" a number of different Web resources. For example, the Web page http://pseudomonas-syringae.org/ contained diverse state-of-the-art datasets related to various *Pseudomonas syringae* pathogenic strains. This page also provided a bridge to other Web resources where additional datasets were collected[1]. After the initial dataset was gathered, filtered, and revised, modelling of the data, and the knowledge within it, was performed using the ontology editor Protégé.

2.1. Design principles and high-level overview of classes

The main goal pursued during the modelling and designing process was to semantically capture as much of the biological knowledge within the data as possible(Figure 1). In particular, a special effort was made to model the *disease triangle* [2]. This idea, one of the cornerstones of plant pathology, asserts that three factors must be present for a disease to occur: a virulent pathogen, a susceptible host, and a propitious environment. Two semantic classes were created to represent these three elements in sufficient richness and precision, namely, the Environmental parameter and the Organism classes. The later class, intended to model phylogenetic taxonomies, is broken into two subclasses that semantically describe either the plant taxonomy, or the pathogenic bacteria's taxonomy. These subclasses are linked to the NCBITaxon_1 class, that incorporates both a bacterial and plant genera hierarchy with their corresponding taxa identification number as provided by the National Center for Biotechnology Information (NCBI). Also, terms such as 'Plant Pathogen', 'Host Plant' or 'Resistant plant' have been strongly axiomatically modelled to deeply capture details within the extracted biological data. Thereafter, by making use of axiomatic reasoning, members of the Host Plant and Resistant Plant subclasses can be automatically inferred (using the FaCT++ and HermiT 1.3.8 reasoners) [[these should be referenced]]. As such, the *reason* that a plant is considered a host, or a bacteria is considered a pathogen, can be determined by exploration of the data, rather than by rote assertion of this role.

The physiological state of a plant can often be inferred visually by observing its expressed phenotype. Thus, significant effort was also made towards modelling plant phenotypes in the PPIO, in the form of several classes specifically created to meticulously represent plant phenotypic traits. Of particular importance are the Phenotype and the 'Phenotypic process' classes. These two classes semantically capture the output of the interaction between the host and the bacteria, which is ultimately represented as a resistance or a susceptibility phenotype.

Finally, the Trait class contains various physiological, biochemical and molecular plant traits. This class was built by importing the Plant Trait Ontology [3] platform [10] into PPIO. The traits described in the different PTO classes are used when describing the effects of a bacterial attack on a susceptible host. This is accomplished by axiomatically relating the Trait class with both Phenotype and 'Phenotypic process' classes within the PPIO.

3. Creation methodology

3.1. URI design

The ontology URI (`http://purl.oclc.org/PPIO`) is HTTP resolvable and permanent (the PURL server redirects to our laboratory's server at `biordf.org`, but could redirect to another location if the ontology is ever migrated to a new location). The identifiers for entities (classes, individuals and object properties) are alphanumeric, with a URI of the type `http://purl.oclc.org/PPIO#PPIO_NNNNNNN`. Every entity has an informative `rdfs:label` annotation. Currently, ontological terms are placed after a

[1] `http://ncppb.fera.defra.gov.uk/`

[2] `http://www.apsnet.org/edcenter/instcomm/TeachingArticles/Pages/DiseaseTriangle.aspx`

[3] `http://www.gramene.org`

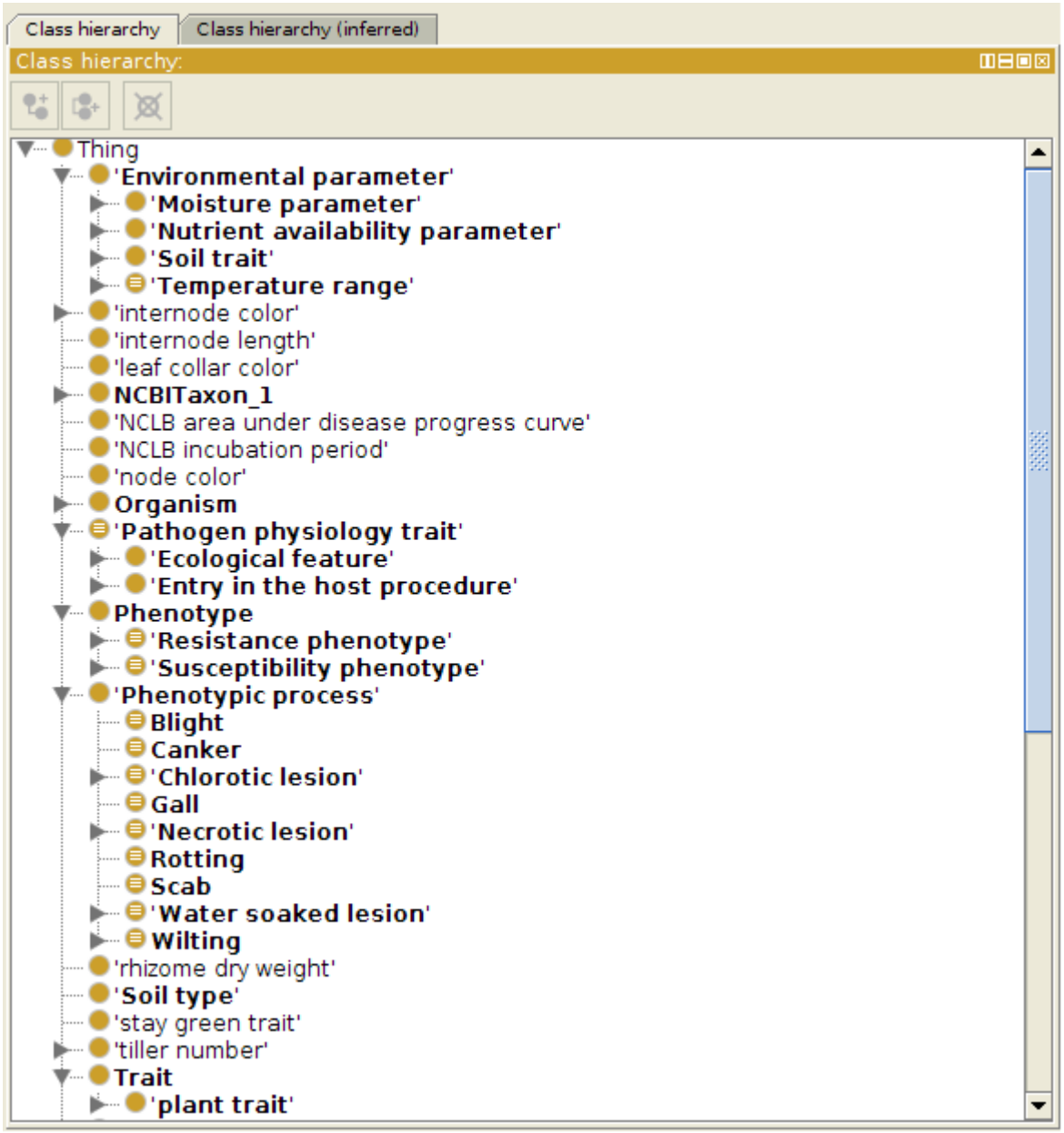

Fig. 1. General view of the PPIO main classes and subclasses. Classes in bold refer to those newly created for building PPIO; the rest of classes are inherited from the Plant Trait Ontology.

"hash" in a common root URI. Due to the small size of the ontology, this does not result in deterimental web-transfer overhead; however since URIs are generated programmatically[4] (see below), if/when the ontology grows into a large-scale Knowledge Base and Linked Data dataset other forms of URIs could be generated that are more "bandwidth-friendly". [[[I wonder if this is worth saying... because the *consequence* of doing this would be that all existing Linked-data statements about the original URI would then be broken, and the community would hate us! ...why are we not using slashed URIs from the get-go?]]]

3.2. Ontology production

The development of PPIO is automated as much as possible, since it is designed to be expanded to adapt to new species/phenotypes as they are described. The core ontological structure is manually constructed. Subsequently, most of the remaining ontological concepts are produced programmatically using the Galaxy platform, a bioinformatics-oriented workflow environment [8]. Within Galaxy, the workflow required to automatically enrich the PPIO is defined once, and then executed for each release. This also allows us to plug PPIO synthesis directly into other Bioinformatics tools, thus enabling and simplifying additional ontological enrichment in the future.

The workflow adds the necessary entities and axioms[5] as follows (Figure 2):

1. The organism taxa hierarchy is produced by the tool NCBITaxonomy2OWL[6]. It retrieves user-defined taxa from the NCBI taxononomy through a BioPortal Web Service [17] and injects these into the PPIO, reproducing the original taxonomical hierarchy (representing each rank-subrank as a simple subsumption relation [14]) and adding each taxon with a resolvable OntoBee[7] URI.
2. Since pathogens in PPIO are modelled as OWL individuals, rather than OWL classes, they cannot be directly related with class hierarchies like the NCBI taxonomy and the symptoms hierarchy. Therefore, PPIO exploits OWL punning[8], where an individual with the same URI as each type-class is generated programmatically for those hierarchies. The linking of pathogens to those hierarchies [[[I don't understand the predicate 'types' in this example...???]]] (*e.g.* `NCBITaxon_552 types Erwinia amylovora`, `NCBITaxon_552 causes symptom Canker`, `NCBITaxon_552 causes symptom Blight`) is done manually. This is achieved by [[[wait.... these two sentences don't make sense... "is done manually -> is achieved by defining two scripts" are contradictory statements]]] defining two Ontology Pre Procesor Language (OPPL)[9] scripts and executing them via OPPL-Galaxy [1] as follows:

```
?x:CLASS,
?y:INDIVIDUAL = create(?x.RENDERING)
SELECT ?x SubClassOf NCBITaxon_1
WHERE ?x != Nothing, ?x != Thing
BEGIN
ADD ?y Type ?x
END;
```

```
?x:CLASS,
?y:INDIVIDUAL = create(?x.RENDERING)
SELECT ?x SubClassOf PPIO_0000069
WHERE ?x != Nothing, ?x != Thing
BEGIN
ADD ?y Type ?x
END;
```

Through our use of OPPL, any complex axiomisation - not only puning - can be defined once and automatically applied every time the workflow is executed.

4. Discussion

Semantic-oriented initiatives like the OBO foundry [15], which includes Gene Ontology (GO) [4], Bio2RDF [2] or the W3C Semantic Web for Health Care and Life Sciences Interest Group[10] have provided copious evidence supporting the successful application of semantics to the problem of automated data integration and exploration. Nevertheless, within the fields of plant

[4] https://github.com/wilkinsonlab/OWLNumericIDGenerator

[5] The workflow can be reproduced at http://biordf.org:8090/u/alejandroriglesias/w/ppio-taxa-punning, and a Galaxy page is available at http://biordf.org:8090/u/alejandroriglesias/p/using-galaxy-tools-in-plant-pathogen-interactions-ontology-ppio

[6] https://github.com/wilkinsonlab/NCBITaxonomy2OWL

[7] http://www.ontobee.org/

[8] http://www.w3.org/TR/owl2-new-features/Punning

[9] http://oppl.sf.net

[10] http://www.w3.org/blog/hcls/

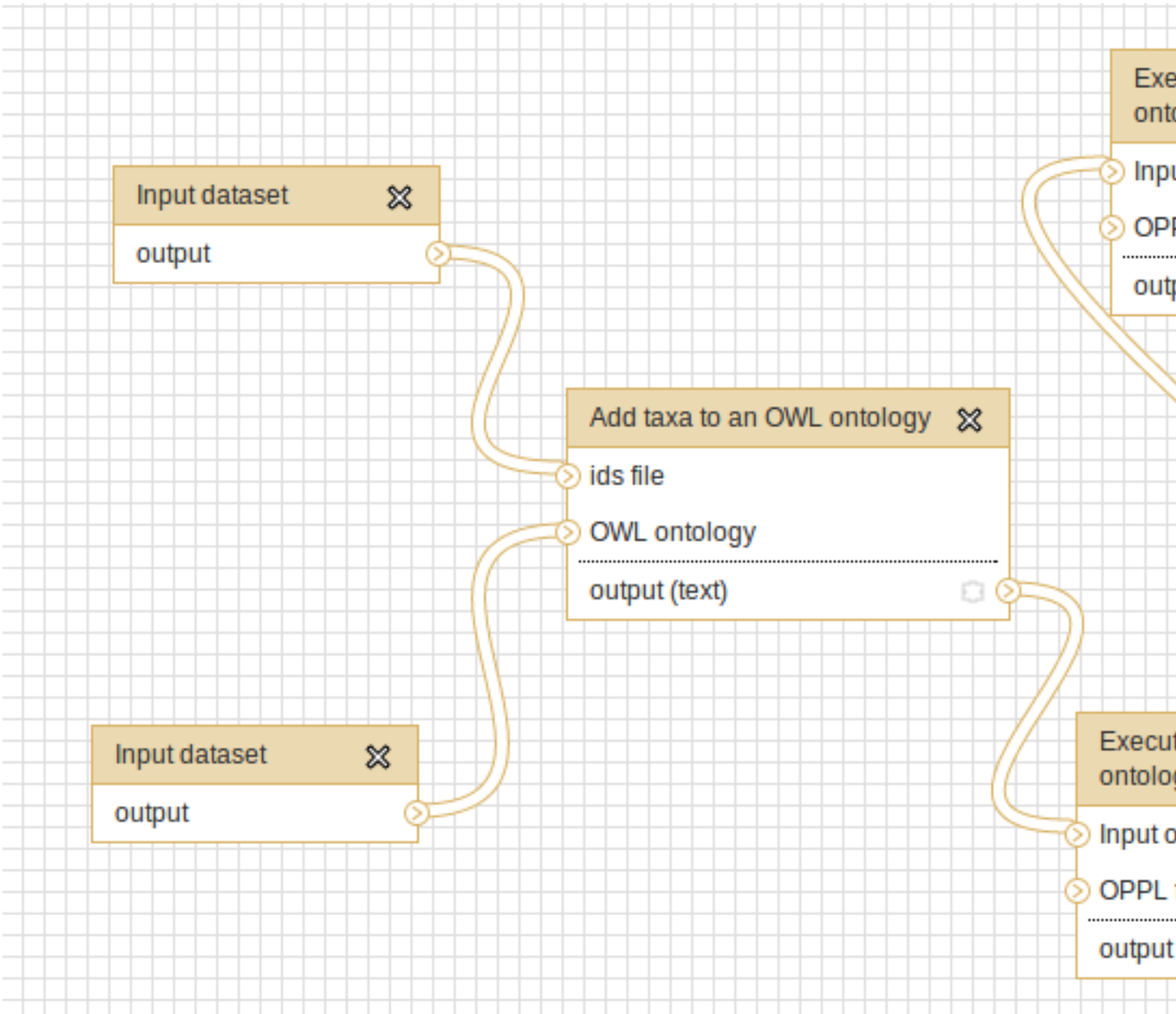

Fig. 2. Galaxy workflow for producing a release of PPIO. In the first step, NCBITaxonomy2OWL is executed; it gets the ontology and a flat file containing the NCBI taxonomy IDs, and it adds them to the ontology. Then two OPPL scripts are executed against the resulting ontology, adding axioms and entities to create

biotechnology and phytopathology, it is surprising how limited the application of these powerful technologies have been, and to how few resources. Some notable exceptions include the Plant Ontology[11], which describes plant anatomy, morphology and developmental stages [9]; and the Plant Disease Ontology (IDOPlant) [16] [5], which, is focused on generically describing plant infectious diseases. Finally, the GO extension for description of the Type III Effectors [11] is perhaps the ontological contribution in the plant pathology and microbiology area most related to PPIO.

Comparison between the IDOPlant and PPIO will reveal that the former constructs a more general model of infection where, for example, plant infectious diseases are described as being caused by either biotic or abiotic agents. The PPIO ontology, in contrast, pursues a knowledge capture strategy focused on more detailed data concerning interactions between plants and their pathogens, and the consequences of those interactions. With this in-mind, the GO extension for type III effectors is designed to capture information about processes at the host-pathogen level, but with a strong emphasis on effector protein data. As such, the PPIO fills a gap in description of pathogenic interactions that does not significantly overlap with (and in fact, explicitly utilizes and extends) other relevant semantic resources. Our objective in the future is to continue integrating data and other knowledge resources, such as the Darwin Core glossary of terms (DwC) [18]. The end-goal is to create a platform that, combined with others, could act as a key component within a diagnosis/prevention/alert system, much like clinical decision support systems in the medical domain. Alone, PPIO acts as a more generalized knowledge base. For example, PPIO will make it possible for users to pose, and answer, questions such as:

1. Is *Solanum lycopersicum* susceptible to the attack of *Pseudomonas syringae* pv. *tomato* DC3000?
2. Does a high humidity favours the development of *Pectobacterium carotovorum* subsp. *carotovorum*?
3. What is the phenotype of the disease produced by *Dickeya dadantii* in *Solanum tuberosum*?
4. What is the host range of the pathogen *Pseudomonas marginalis* pv. *marginalis*?

Knowledge acquisition is the process of converting knowledge from unstructured sources into a format that is more rigorously processable. As datasets become larger, and as expert knowledge continues to be broadly dispersed, it becomes increasingly necessary to automate the capture of relevant knowledge, and optimally, to automate the addition of this knowledge into a rich ontological context. However, while the PPIO represents a framework for capturing rich plant/pathogen interaction data in a machine-processable manner, we continue to rely on field experts to ensure the accuracy of the data content captured within it. To this end, we are pursuing a knowledge capture project specifically aimed at accurately collecting relevant, manually-verified data, *en masse* that will ultimately populate the PPIO knowledgebase. We hope, thereby, to achieve the objective of making PPIO an essential bioinformatics tool for the plant-pathogen community.

[11] `tp://www.plantontology.org/`

Acknowledgements

Alejandro Rodríguez Iglesias and Alejandro Rodríguez González are funded by the Isaac Peral Programme. Mark D. Wilkinson and Mikel Egaña Aranguren are funded by the Marie Curie-COFUND Programme (FP7) of the EU.

References

[1] M. E. Aranguren, J. T. F. Breis, E. Antezana, C. Mungall, A. R. González, and M. Wilkinson. OPPL-Galaxy, a Galaxy tool for enhancing ontology exploitation as part of bioinformatics workflows. *Journal of Biomedical Semantics*, 4(1):2+, 2013.

[2] F. Belleau, M.-A. Nolin, N. Tourigny, P. Rigault, and J. Morissette. Bio2rdf: Towards a mashup to build bioinformatics knowledge systems. *Journal of Biomedical Informatics*, 41(5):706–716, 2008.

[3] C. T. Bull, S. H. D. Boer, T. P. Denny, G. Firrao, M. F.-l. Saux, G. S. Saddler, M. Scortichini, D. E. Stead, and Y. Takikawa. LETTER TO THE EDITOR COMPREHENSIVE LIST OF NAMES OF PLANT PATHOGENIC BACTERIA , 1980-2007. 92:551–592, 2010.

[4] G. O. Consortium. The Gene Ontology in 2010: extensions and refinements. *Nucleic acids research*, 38(Database issue):D331–5, 2010.

[5] L. G. Cowell and B. Smith. Infectious disease ontology. *Infectious Disease Informatics*, pages 373–395, 2010.

[6] P. J. G. M. de Wit. How plants recognize pathogens and defend themselves. *Cellular and molecular life sciences : CMLS*, 64(21):2726–32, 2007.

[7] P. Dodds and J. Rathjen. Plant immunity: towards an integrated view of plant-pathogen interactions. *Nature Review Genetics*, 11:539–548, 2010.

[8] J. Goecks, A. Nekrutenko, J. Taylor, and Galaxy Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome biology*, 11(8):R86+, 2010.

[9] P. Jaiswal, S. Avraham, K. Ilic, E. A. Kellogg, S. McCouch, A. Pujar, L. Reiser, S. Y. Rhee, M. M. Sachs, M. Schaeffer, L. Stein, P. Stevens, L. Vincent, D. Ware, and F. Zapata. Plant ontology (po): a controlled vocabulary of plant structures and growth stages. *Comparative and Functional Genomics*, 6(7-8):388–397, 2005.

[10] P. Jaiswal, D. Ware, J. Ni, K. Chang, W. Zhao, S. Schmidt, X. Pan, K. Clark, L. Teytelman, S. Cartinhour, L. Stein, and S. McCouch. Gramene: development and integration of trait and gene ontologies for rice. *Comparative and Functional Genomics*, 3(2):132–136, 2002.

[11] M. Lindeberg and A. Collmer. Gene Ontology for type III effectors: capturing processes at the host-pathogen interface. *Trends in microbiology*, 17(7):304–11, 2009.

[12] J. Mansfield, S. Genin, S. Magori, V. Citovsky, M. Sriariyanum, P. Ronald, and et al. Top 10 plant pathogenic bacteria in molecular plant pathology. *Molecular Plant Pathology*, 13(6):614–629, 2012.

[13] E. Montesinos. Pathogenic plant-microbe interactions. What we know and how we benefit. *International microbiology : the official journal of the Spanish Society for Microbiology*, 3(2):69–70, 2000.

[14] S. Schulz, H. Stenzhorn, and M. Boeker. The ontology of biological taxa. *Bioinformatics*, 24(13):i313–321, July 2008.

[15] B. Smith, M. Ashburner, C. Rosse, J. Bard, and W. B. et al. The obo foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotech*, 25(11):1251–1255, 2007.

[16] R. Walls, B. Smith, J. Elser, A. Goldfain, D. W. Stevenson, and P. Jaiswal. A plant disease extension of the Infectious Disease Ontology. *In Ronald Cornet and Robert Stevens, editors, ICBO*, pages 1–5, 2012.

[17] P. L. Whetzel, N. F. Noy, N. H. Shah, P. R. Alexander, C. Nyulas, T. Tudorache, and M. A. Musen. BioPortal: enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications. *Nucleic acids research*, 39(Web Server issue):W541–W545, 2011.

[18] A. J. Wieczorek, O. Bánki, S. Blum, J. Deck, M. Döring, G. Dröge, P. Goldstein, P. Leary, L. Krishtalka, E. O. Tuama, and J. Robert. Meeting Report : GBIF hackathon-workshop on Darwin Core and sample data (22-24 May 2013). (May), 2013.