Semantic Web 0 (0) 1
IOS Press

From pure phytopathology to the Semantic Web world: the Plant-Pathogen Interactions Ontology (PPIO)

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

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

Abstract. The scientific relevance of the studies based on plant-pathogenic bacteria interactions is undeniable. This is also important from the economic point of view of some plants species, and, of course, there is also a biodiversity component behind these studies. The implementation of the tools the Semantic Web offers into this area of knowledge has not been completely carried out. We present here the Plant-Pathogen Interactions Ontology (PPIO), whose axiomatic models allow the integration and inference of plant-pathogenic bacteria interactions datasets in an automated manner.

Keywords: Plant pahogenic bacteria, Ontologies, Semantic Web, PPIO

1. Introduction

Plants can be susceptible to the attack of different pathogenic bacterial genera [12]. If the different levels of the plant defense barriers are overcomed, the infection process can ultimately lead to the death of the plant. This phenomenon affects crop yield, and as a consequence there is also an impact on the economy around this species [13]. Being able to explore new data sources data regarding plant-pathogenic bacteria interactions is an essential step, and will ultimately lead to a better preservation of worldwide crops. Hundreds of publications that focus their interest on unveiling the mechanisms of pathogenic bacteria interactions with their hosts reflect the biological significance of this area of research [6] [7].

The field of plant-pathogenic bacteria is a good example of biodiversity richness [3]. Traditionally, *Agrobacterium, Erwinia, Pseudomonas and Xanthomonas* were considered the four main plant pathogenic genera. With the rapid improvement of the research experimental approaches, the number of known plant pathogenic bacteria genera has increased up to 30. This has resulted in the generation of big amounts of unexplored biological data. The effectiveness of the use of Semantic technologies to manage with large sets of data has been already proved in other areas of life sciences, as discussed later. However, these tools that the Semantic Web offers have not been extensively applied to the knowledge domain of plant-pathogenic bacteria to date. Here we present the Plant Pathogen Interactions Ontology (PPIO), an ontology developed to collect data from the plant-pathogenic bacteria interactions domain. The main goal of PPIO is to serve as a reference for expert plant pathologists, providing the knowledge necessary to provide assistance in the interpretation of the phenotypic responses that result from these biological interactions.

2. Modelling

The initial data necessary to start building PPIO was collected manually by consulting a number of different

[*] Corresponding author. Email:alejandroriglesias@gmail.com

expert web resources. The web page http://pseudomonas-syringae.org/ contained diverse state-of-art datasets related with various *Pseudomonas syringae* pathogenic strains. This page was a bridge to other web services where more datasets were collected [1]. After an initial collection of data was gathered and revised, the modelling of these datasets was performed using the ontology editor Protégé.

2.1. Desing principles

The main goal pursued during the modelling and designing process was to semantically capture as many biological data existing as possible, and special effort has been done in modelling the *disease triangle* [2]. This term, one of the essentials milestones in 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 classes have been created to represent these three elements in a truthful and precise manner, the Environmental parameter and the Organism classes. This later class contains two subclasses that semantically describe both plant and pathogenic bacteria; these subclasses are linked to the NCBITaxon_1 class, that incorporates both a bacterial and plant genera hierarchy with their corresponding taxa identifier. The connection between the three components of the disease triangle and the description of each one in the ontology is vital for accurately expressing this biological information into ontology language. To this end, terms such as 'Plant Pathogen', 'Host Plant' or 'Resistant plant' have been strongly axiomatically modelled to assure a trustworthy capture of the extracted biological data. By making a good use of the reasoning power this technology offers, the members of the Host Plant and Resistant Plant subclasses were inferred after the reasoning process took place (using the FaCT++ and HermiT 1.3.8 reasoners).

Physiological state of plants can be inferred visually by observing different phenotypes. Thus, significant endeavour has also been done in modelling plant phenotypic representations in PPIO. Therefore, a number of classes have been specifically created to meticulously represent plant phenotypic traits. Specially important are the Phenotype and the 'Phenotypic process' classes. These two classes semantically illustrate the output of the interaction between the host and the bacteria, which is ultimately represented as a resistance or a suscepibility phenotype expressed in the plant.

The Trait class contains various physiological, biochemical and molecular plant traits. To be able to work with the classes that contain these traits, the Plant Trait Ontology [3] platform [10] has been imported and integrated into PPIO, converting this ontology into an archetype of the Semantic Web technologies capabilities for data integration. The traits described in the different PTO classes can be affected if a bacterial attack takes place, and this has also been semantically illustrated in PPIO by axiomatically relating the Trait class with both Phenotype and 'Phenotypic process' classes.

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 current server at `biordf.org`). The identifiers for entities (classes, individuals and object properties) are alphanumeric, with a URI of the type `http://purl.oclc.org/PPIO#PPIO_NNNNNNN`, and every entity has an informative `rdfs:label` annotation. Currently hash URIs are used due to the small size of the ontology, but since URIs are generated programmatically[4], when the ontology grows into a Knowledge Base or Linked Data dataset (see section 4) slash URIs can be generated.

3.2. Ontology production

The development of PPIO is automated as much as possible. Once the main structure is set, most of the remaining parts are produced programmatically using the Galaxy platform, a bioinformatics-oriented workflow environment [8]. By using Galaxy, the specific workflow we need is defined once and executed for each release; also, we can plug PPIO directly with other Bioinformatics tools.

The workflow adds the necessary entities and axioms[5] (Figure 1):

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

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

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

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

[5] The workflow can be reproduced at `http://biordf.org:8090/u/alejandroriglesias/w/ppio-taxa-punning`

1. The organism taxa hierarchy is produced by the tool NCBITaxonomy2OWL[6]: it gets the user-defined taxa from the NCBI taxononomy through a BioPortal Web Service [17] and injects them in 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 individuals, they cannot be directly related with class hierarchies like the NCBI taxonomy and the symptoms hierarchy. Therefore, PPIO exploits OWL punning[8] and an individual with the same URI as each type class is generated programmatically (see bellow) for those hierarchies: the linking of pathogens to those hierarchies (*e.g.* `NCBITaxon_552 types Erwinia amylovora`, `NCBITaxon_552 causes symptom Canker`, `NCBITaxon_552 causes symptom Blight`) is done manually. This is achieved by defining two Ontology Pre Procesor Language (OPPL)[9] scripts and executing them via OPPL-Galaxy [1][10]:

```
?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;
```

4. Discussion

Semantic-oriented platforms 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[11] are excellent paradigms of the success in using semantics to assist in the integration of automated data. Nevertheless, taking a look at the fields of plant biotechnology and phytopathology, it is surprising to notice that these technologies have been applied to a limited number of domain resources. There are some precedents prior to the ontology discussed here, such as the Plant Ontology[12], which describes plant anatomy, morphology and developmental stages [9]. The Plant Disease Ontology (IDOPlant) [16] [5], on the other hand, is focused on generically describing plant infectious diseases. Finally, the GO extension for description of the Type III Effectors [11] is maybe the ontological contribution in the plant pathology and microbiology area more related to PPIO.201

In a comparison between the IDOPlant and PPIO, a more generalistic approach of data modelling can be appreciated in the case of the first platform, which describes plant infectious diseases caused by either biotic or abiotic agents. The ontology reported here pursues a knowledge capture strategy focused on data that concerns plant-pathogenic bacteria interactions, in order to be able to properly represent this area of knowledge. On the other hand, although the GO extension for the type III effectors is built for capturing processes at the host-pathogen level, effector proteins data capture is emphasized. So, the principal reason for building this ontology is to ensure the semantic description of a domain not entirely represented by other resources. Of course, PPIO has been designed to complement these previous ontologies by introducing accurate biological information. It is our objective to continuing the process of data integration not only importing related ontologies, but other resources such as the Darwin Core glossary of terms (DwC) [18]. The final goal of this initiative is to use this platform, combined with others, as a diagnosis/prevention/alert system. PPIO will make it possible for users to pose, and answer, questions like the following ones and obtain a meaningful answer:

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*?

[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] Since we are using OPPL, any complex axiomisation -not only puning- can be defined once and automatically applied -expanded to different parts of the ontology- every time the workflow is executed.

[11] `http://www.w3.org/blog/hcls/`

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

Fig. 1. 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

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 based on the process of transcripting the knowledge from unstructured sources into a format that is machine-readable and useful. This approach can report a great benefit if big amounts of datasets are required in order to populate ontologies. One of the main ideas behind this is that the participation of field experts should ensure the fiability of the data content captured. At the moment, a knowledge capture project is now being developed in our laboratory, and it is our thought that it will aid in the process of collecting data that will ultimately populate PPIO with trustworthy and scientifically relevant information. All these steps will help to our objective of converting PPIO into a basic and essential bioinformatic tool for scientific community in the area plant pathogens.

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.