A plant disease extension of the Infectious Disease Ontology

Ramona Walls[1], Barry Smith[2], Justin Elser[3], Albert Goldfain[4]

Dennis W. Stevenson[1], Pankaj Jaiswal[3]

1. New York Botanical Garden, Bronx, NY, USA,
2. Department of Philosophy, University at Buffalo, Buffalo, NY, USA,
3. Department of Botany and Plant Pathology, Oregon State University, Corvallis, OR, USA,
4. Computer Science Department, Blue Highway, Inc., Syracuse, NY, USA

ABSTRACT

Plants from a handful of species provide the primary source of food for all people, yet this source is vulnerable to multiple stressors, such as disease, drought, and nutrient deficiency. With rapid population growth and climate uncertainty, the need to produce crops that can tolerate or resist plant stressors is more crucial than ever. Traditional plant breeding methods may not be sufficient to overcome this challenge, and methods such as high-throughput sequencing and automated scoring of phenotypes can provide significant new insights. Ontologies are essential tools for accessing and analysing the large quantities of data that come with these newer methods. As part of a larger project to develop ontologies that describe plant phenotypes and stresses, we are developing a plant disease extension of the Infectious Disease Ontology (IDOPlant). The IDOPlant is envisioned as a reference ontology designed to cover any plant infectious disease. In addition to novel terms for infectious diseases, IDOPlant includes terms imported from other ontologies that describe plants, pathogens, and vectors, the geographic location and ecology of diseases and hosts, and molecular functions and interactions of hosts and pathogens. To encompass this range of data, we are suggesting in-house ontology development complemented with reuse of terms from orthogonal ontologies developed as part of the Open Biomedical Ontologies (OBO) Foundry. The study of plant diseases provides an example of how an ontological framework can be used to model complex biological phenomena such as plant disease, and how plant infectious diseases differ from, and are similar to, infectious diseases in other organism.

1 INTRODUCTION

Plants are the primary food source on which almost every other organism on earth depends, either directly or indirectly, and six plant species – wheat, rice, corn, potato, sweet potato, and cassava – provide 80% of calories consumed by humans worldwide (FAO, 2012; Goudie & Cuff, 2001). It is imperative to develop higher-yielding crop varieties to support the growing human population. This can be done in two primary ways, (1) by increasing, e.g., the number or size of grains on a cereal plant or tubers on a potato plant, and (2) by reducing losses due to diseases and pests. Pre-harvest disease and pest damage in the eight most important food and cash crops in the world account for ~42% of attainable production, and infectious plant diseases also threaten plant conservation and human health (Anderson et al., 2004).

Many challenges in plant pathology (the study of plant diseases) can potentially be met through advances in methods such as high-throughput sequencing and automated scoring of phenotypes (Studholme et al., 2011; Furbank & Tester, 2011). Complete genome sequences already exist for 25 green plant species, of which 17 are agriculturally important (Anon, 2012), along with expression sequence tags (EST), unigene, mutant phenotype, and other data sets for hundreds of plant species. Additionally, a vast quantity of information on plant diseases is available in resources like manuals, textbooks, extension program highlights, and crop management databases, but almost always in natural language form. Access to and analysis of the growing quantities of genomic, phenomic, and free-text data can be greatly facilitated when data are annotated using ontologies. The development of ontologies can also foster consistency in the description of plant diseases, including aspects such as environmental factors, areas of endemism, phenotypes associated with diseases, and development stages of both plants and pathogens. Finally, the standardization and reasoning power provided by using ontologies enhances data sharing among biomedical researchers, allowing the results of research in plant pathology to be translated into applications for human or other animal diseases, and vice versa.

A plant disease is traditionally defined as a deviation from normal physiological functioning that is harmful to a plant (Manners, 1993). Biotic factors or stressors such as pests or pathogens and abiotic factors such as low temperature, air pollution, or nutrient deficiency, may cause plant diseases. Infectious plant diseases are caused by pathogens, such as fungi, bacteria, and viruses. As part of a larger project to develop ontologies that describe both biotic and abiotic plant stresses, we are developing a plant disease extension (IDOPlant) of the Infectious Disease Ontology (IDO) (Cowell & Smith, 2010) as a reference ontology for plant disease. The goals are to provide plant scientists with the means to identify genomic and genetic signatures of host-pathogen interactions, resistance, or susceptibility, and to help agronomists and farmers by developing tools to identify disease phenotypes and gather epidemiological statistics.

IDOPlant will integrate and interoperate with member and candidate ontologies of the Open Biomedical Ontologies (OBO) Foundry (Table 1), such as: the Plant Ontology (PO; describes the *plant structures* and the *development stages* at which infections happen or signs of disease are observed), the Plant Trait Ontology (TO; describes phenotypes or entities that are evaluated in plants, such as *leaf color* or *grain yield*), and the Gene Ontology (GO; describes

* To whom correspondence should be addressed:
jaiswalp@science.oregonstate.edu

ID	Ontology Name	Domain
PO	Plant Ontology[1]	plant structures and development stages
GO	Gene Ontology[2]	biological processes, sub-cellular components, molecular function
TO	Trait Ontology[3]	plant traits
PATO	Phenotypic Quality Ontology[4]	phenotypic qualities
OBI	Ontology for Biomedical Investigations[5]	protocols, instrumentation, materials, data, types of analysis
ENVO	Environment Ontology[6]	environmental features and habitats
ChEBI	Chemical Entities of Biological Interest[7]	chemical entities
GAZ	Gazetteer[8]	geographical information
NCBItaxon	NCBI Taxonomy Classification[9]	taxonomic classification of living organisms

Table 1. Some of the external ontologies needed to describe plant diseases. References: 1. Ilic et al., 2007, 2. Gene Ontology Consortium, 2010, 3. Jaiswal, 2011, 4. Mungall et al., 2010, 5. http://obi-ontology.org/page/Main_Page, 6. http://http://www.environmentontolo gy.org/, 7. Degtyarenko et al. 2007, 8. http://bioportal.bioontolgy.org/ontologies/1397, 9 http://www.obofoundry.org/wiki/index.php/NCBI Taxon:Main_Page.

the *molecular functions* of interacting genes from host and pathogen as well as *biological processes* involving either host, pathogen, or both). The *multi-organism process* branch of the GO, developed as part of the PAMGO project (Giglio et al., 2009), is especially relevant to the IDOPlant. To ensure compatibility with research on non–plant diseases, the IDOPlant is created by downward population from the upper-level terms of the IDO. The IDOPlant differs from existing IDO extensions, because the latter focus on specific diseases or pathogens, like *Malaria* or *Brucellosis*, that affect human or other animal health (Lin et al., 2011; Topalis et al., 2010). IDOPlant, in contrast, is designed to encompass any plant infectious disease. Furthermore, the IDOPlant is being developed as part of the larger Plant Phenotype and Stress Ontology Project, which is not limited to infectious diseases but encompasses any plant stress. Our approach calls for a multi-pronged strategy that includes creating new terms, as well as importing terms from, and building links to, other ontologies.

The study of plant diseases provides an excellent example of how the framework of the OBO Foundry (Smith et al. 2007) allows us to describe complex biological phenomena using terms from multiple ontologies. By constructing the IDOPlant within the OBO framework, we eliminate redundant efforts, have a head start in ontology term development, and yield outcomes compatible with databases that already annotate their data using OBO Foundry ontologies, such as the Arabidopsis Information Resource (TAIR) (Swarbreck et al., 2008), Gramene (Youens-Clark et al., 2011; Jaiswal, 2011) and Uniprot (The UniProt Consortium, 2010). In this paper, we describe our plans for the overall structure of the IDOPlant, provide an example of how to model plant disease data, and discuss the types of data that can be annotated with the IDOPlant.

2 METHODS

Throughout this paper, words in italics are ontology terms, e.g., *pathogen*. If the source ontology is not evident from the context, then we prefix with the ontology ID, as in: IDO:*pathogen*. The IDOPlant and the Plant Phenotype and Stress Ontology are being constructed in web ontology language (OWL), using the Protégé 4.1 software (http://protege.stanford.edu) The annotation standard format will follow the GO and PO model with the GAF2.0 format (Gene Ontology, 2012).

We began by reviewing whether the terms in the IDO were adequately structured for describing plant infectious diseases, including discussion with the developers of IDO and the Ontology for General Medical Science (OGMS; http://code.google.com/p/ogms/). Next, following the strategy used in other IDO extensions, we created terms for the IDOPlant, such as *plant infectious disease*, specific to the needs of plant pathology. More specific terms, such as *rice bacterial blight disease* were added as an example of how to model a specific disease and to provide terms to be used in annotating existing gene expression data available through Gramene (http://www.gramene.org). Textual definitions and relationships among terms are drawn from plant pathology textbooks or journal articles, and are reviewed by plant disease experts.

Logical definitions for IDOPlant terms are being constructed in OWL. Many of the terms needed for logical definitions already exist in other ontologies. To access these terms, we could import entire ontologies into the IDOPlant, but this would result in many unnecessary terms and may cause problems if the resultant ontology is too large. Importing a selected subset of terms creates problems as well. If we import individual terms from external ontologies, then we lose the ontology structure they are associated with and the reasoning power that comes with it. If we import selected terms through the MIREOT process (Courtot et al., 2009), which imports the minimum information to reference an external ontology, we have to update the IDOPlant whenever there is a change to the source ontology.

To cope with these issues, we use a multi-pronged strategy that includes directly importing some terms and building bridge files to link to external ontologies.

- Terms specific to plant diseases are added to the IDOPlant and assigned unique IDOPlant IDs, e.g., IDOPlant:#######.
- Terms falling near the bottom of the IDOPlant hierarchy that are drawn from ontologies from which only a few terms are needed are imported as single terms, using the original ontology ID. When appropriate, the MIREOT method is used.
- Content treated in ontologies from which many terms are needed are accessed by simultaneously loading multiple ontologies and creating relations among them using

bridge files (Mungall et al., 2010). This applies specifically to the three main ontologies (PO, TO, and PATO) whose terms are needed to describe plant stresses. Users will be required to open the entire suite of these ontologies when annotating data with the IDOPlant.

- Taxonomic entities require special treatment, because we will ultimately need to import many terms for plant species, disease organisms, and vector species. However, the NCBItaxon ontology is very large and can be impractical to work with when loaded. Therefore, we will manually import the necessary taxa into the IDOPlant from either NCBItaxon or uBio (http://ubio.org).
- In the event that a term imported into the IDOPlant is made obsolete in the source ontology, we will replace the term either with the term suggested by the source ontology or with a new term created for the IDOPlant.

3 RESULTS AND DISCUSSION

Researchers should contact the curators before using the IDOPlant, because it is under active development. A draft is available at http://purl.obolibrary.org/obo/idoplant.owl.

3.1 Using IDO for plant infectious diseases

Our review of the IDO suggests that it is generally appropriate as a foundation for the description of plant diseases. The IDO is rooted in the Basic Formal Ontology (BFO) (Arp & Smith, 2008) and in the OGMS, which increases compatibility with other OBO Foundry ontologies and helps to ensure logically consistent use of type-subtype relations. For example, *IDO:pathogen* cannot be classified as a subtype of *IDO:process of establishing an infection*, because they belong to disjoint super-classes (BFO:*continuant* and BFO:*occurrent*, respectively).

The IDO consists primarily of terms specific to *infectious disease*, together with relevant terms imported from other ontologies, such as *organism* from OBI; *disease*, *disorder*, and *disease course* from OGMS; *habitat* from ENVO; *macromolecular complex*, *reproduction*, and *entry into host* from GO; *bacterium* and *virus* from NCBItaxon; and *molecular entity* from ChEBI. The IDO has created many new terms, such as *resistance to drug*, *infectious agent*, and *infectious disease epidemic*. The bulk of the unique IDO terms can be used for the IDOPlant without modification. For example IDO:*infectious disease* is defined as "A disease whose physical basis is an infectious disorder". This in turn is based on the OGMS definition of disease: "A disposition (i) to undergo pathological processes that (ii) exists in an organism because of one or more disorders in that organism" (Scheuermann et al., 2009). Although the wording of definitions such as this may be unfamiliar to plant pathologists, the meaning is consistent with traditional treatments of plant disease (e.g., Manners, 1993).

IDO terms such as *transition to clinical abnormality* or *subclinical infection* required careful consideration, because the word "clinical" is not commonly used for plants. We decided that the meaning of their definitions was appropriate for plants, despite the names. For example, a feature of an organism is clinically abnormal when it: "(1) is not part of the life plan for an organism of the relevant type … (2) is causally linked to an elevated risk either of pain or other feelings of illness, or of death or dysfunction, and (3) is such that the elevated risk exceeds a certain threshold level" (Scheuermann et al., 2009). All three conditions can be met in plants. Although we cannot know if plants experience pain or feelings of illness, we can assess death or dysfunction in plants.

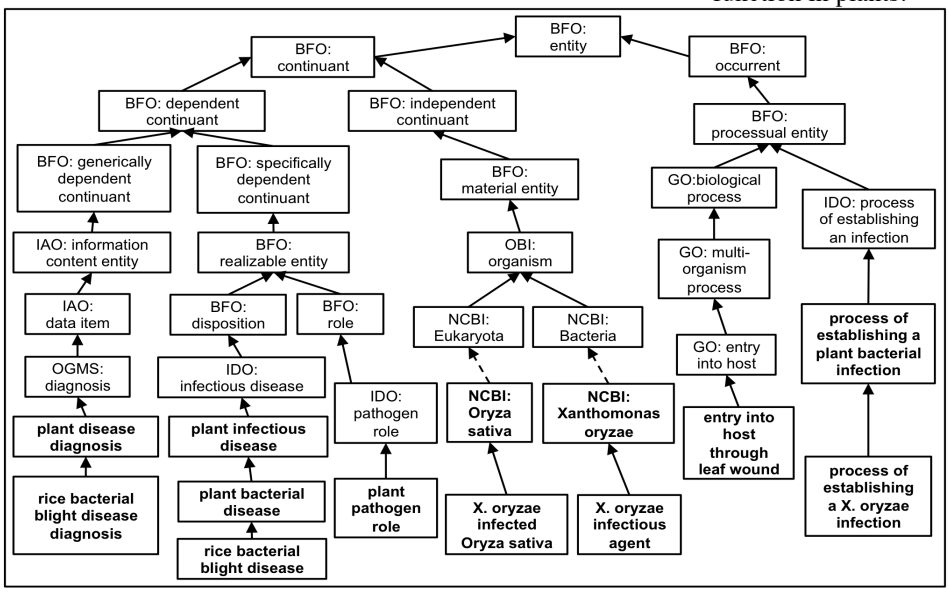

Fig. 1. Selected terms from the upper-level type-subtype hierarchy of the IDOPlant, with top-level terms imported form the IDO (including terms imported to the IDO from BFO and other ontologies) and lower-level terms that were added as part of the IDOPlant (in bold). All arrows represent *is_a* relations. Dashed arrows indicate several skipped intermediate steps in the ontology hierarchy.

Another potential limitation of the IDO for use in plant science is the meaning of terms from the OGMS that were defined within the scope of clinical encounters involving humans. In particular, the definition of *symptom* from OGMS requires a host of a type that can report its experiences, and so is restricted to sentient hosts. In plant pathology, "symptom" is commonly used to describe the phenotypes that are associated with a plant disease. The phenotypes are independent of the disease and the same phenotype or "symptom" may be associated with many different diseases. Furthermore, diagnosis generally depends on a collection of phenotypes, and not every instance of a

disease will display every phenotype that is typical of the disease. Rather than use the OGMS definition of *symptom*, we developed a new term for the IDOPlant:

> *plant disease symptom* =def. A feature of a plant that is of the type that can be hypothesized to be involved in the realization of a plant disease.
>
> *Comment*: Features include phenotypes such *pale yellow leaf color*, processes such as *sudden wilting*, and independent continuants such as *leaf lesion*.

The terms *plant disease symptoms* already exist in other ontologies (primarily the TO), and will be linked to *plant diseases* using the relation *has_plant_disease_symptom* (see section 3.3).

3.2 Terms from external ontologies

The study of plant diseases encompasses many domains. In addition to IDO terms common to all infectious diseases, like *pathogen* or *resistance*, the IDOPlant needs terms to describe the taxonomy of host plants, pathogens, and vectors, genomic and genetic data, the geographic location and ecology of diseases and hosts, plant and fungal anatomy, plant and pathogen development, biological processes, and molecular functions. To encompass this range, the IDOPlant is not only creating new ontology terms specific to its domain, but also integrating and linking to existing terms from multiple sources (Table 1). Whenever possible, existing ontology terms are being used to create logical definitions for IDOPlant terms. For example, *rice bacterial leaf blight* is defined as "A *bacterial blight disease* (in IDOPlant), that has as infectious agent *Xanthomonas oryzae* (from NCBItaxon)" (fig. 3). Terms from external ontologies will also be used for relations such as *rice bacterial leaf blight disease has_plant_disease_symptom pale yellow leaf color* (from TO). Logical definitions allow us to use automated reasoners to ensure that the ontology hierarchy is sound and to infer sub-types and relations implied by the definitions. These can then be added to the ontology if correct or eliminated if incorrect or redundant (Meehan et al., 2011).

3.3 IDOPlant relations

The IDO imports the Relation Ontology (RO) (Smith et al., 2005), which includes basic relations such as *SubClassOf* (*is_a*), *part_of*, *participates_in*, *inheres_in*, *bearer_of*, *has_disposition, has_role,* and *has_function*. In addition, we plan to incorporate several new relations:

has_material_basis: This relation is under development by the OGMS and will be added to the BFO. It is used to indicate the material basis of a *disease*. For infectious diseases, we use the *has_infectious_agent* relation.

has_infectious_agent: This relation, which is under consideration by the IDO, is used to indicate the material basis of an *infectious disease*, e.g., *rice bacterial leaf blight disease has_infectious_agent Xanthamonas oryzae*.

In addition we are developing the following for IDOPlant:

has_plant_disease_symptom: This relation is used to indicate a phenotype, process, or independent continuant that is evaluated to diagnose a disease. For example, "*rice bacterial leaf blight disease has_plant_disease_symptom leaf color pale yellow"* means that *pale yellow leaf color* is a *plant disease symptom* (see above) of *rice bacterial leaf blight disease*, but it does not mean that every instance of *rice bacterial leaf blight disease* has pale yellow leaves.

3.4 Modeling disease in the IDOPlant

Much of the information available on plant diseases is in a natural language or free text form, such as: "Bacterial leaf blight disease of rice is caused by *Xanthomonas oryzae*. It produces pale yellow leaves in mature plants. In one report the pathogen and the disease were reported in the Northern Territory of Australia." Using ontologies to process such descriptions in a standardized form makes them comprehensible to computers and reasoners. For example, the description above could be converted (using natural language processors or other mechanisms) to:

```
disease: rice bacterial leaf blight dis-
ease | host species: Oryza sativa (rice)
| caused by: Xanthomonas oryzae | has
symptom: pale yellow leaves | reported
in: Northern Territory
```

This standardized text could then be made even more powerful using ontology terms and relations (Fig. 3).

3.5 Integrating data into the IDOPlant

The current situation in the plant disease research community is similar to that in the animal community when the GO, MeSH (Savage, 2000), and CARO (Haendel et al., 2007)

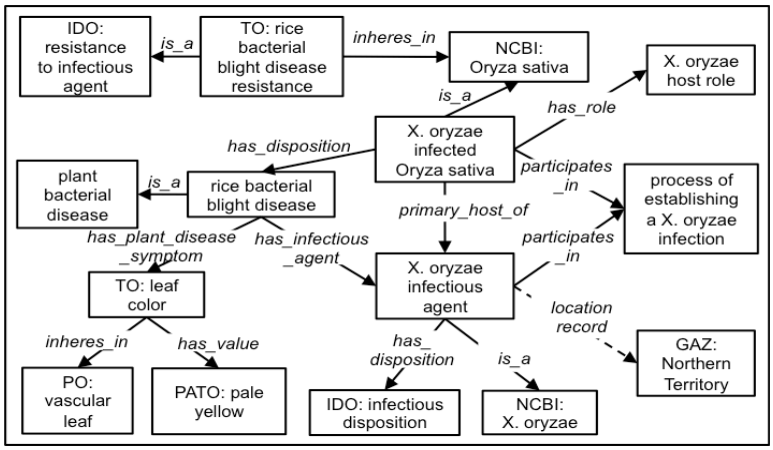

Fig. 3. Some of the terms and relations needed to model *rice bacterial blight disease* in the IDOPlant. Following the IDO, a disease is treated as a disposition of an infected organism, which has a particular infectious agent. The IDOPlant can also be used to define terms in the TO, such as *rice bacterial blight disease resistance*, which is a *resistance to infectious disease* that inheres in *Oryza sativa*.

projects were being initiated: A number of organism-specific databases are faced with large amounts of molecular, germplasm (stock), genotype, and phenotype data associated with function, phenotype, or environment. The sharing of the task of building a set of controlled vocabularies such as GO and PO has helped enormously to address the needs of multiple individual databases. The IDOPlant controlled vocabulary for plant infectious diseases will allow database curators to store and retrieve the results of experiments related to diseases, including quantitative trait loci, pathogen and host germplasm descriptions, microarray expression studies, gene knockouts, reporter gene expression patterns, and gene-gene interactions from host and pathogen. The Plant Phenotype and Stress Ontology Project aims to overcome the obstacles in annotating data for complex biological concepts that span multiple ontologies by developing both the ontology terms and the software tools needed to annotate data from all aspects of plant diseases.

To annotate plant infectious disease description data, the IDOPlant is reaching out to resources such as the Food and Agriculture Administration's AGROVOC (http://aims.fao.org/website/AGROVOC-Thesaurus/sub) and the International Rice Research Institute (http://www.knowledgebank.irri.org/rice.htm). These resources will enrich the IDOPlant by providing a wealth of information that can be incorporated into the ontology and by identifying gaps and errors. The IDOPlant can benefit these organizations by making their content more easily accessible to semantic applications.

4 CONCLUSIONS

As the growing human population and climate change place even more uncertainty on food supply, the need to understand the linkages between plant disease, environment, and yield is greater than ever. The IDOPlant and the Plant Phenotype and Stress Ontology Project can contribute to this challenge by making data on plant diseases more accessible. We are taking advantage of the interoperability of OBO Foundry ontologies to leverage existing terms to enhance the new IDOPlant extension, simultaneously enriching all ontologies involved by filling in terms needed for logical definitions. By expanding the core terms of the IDO to plants, we can learn how plant diseases differ from, and are similar to, infectious diseases in general.

ACKNOWLEDGEMENTS

RW, JE, DWS, and PJ are supported by NSF-IOS: 0822201 to the Plant Ontology Project. BS is supported by NIH:U54 HG004028 (National Center for Biomedical Ontology) and NIH / NIAID R01 AI 77706-01 (Immune System Biological Networks). Thanks to Lindsay Cowell, Laurel Cooper, Robert Hoehndorf, and three anonymous reviewers for comments on earlier versions of this manuscript.

REFERENCES

Anderson, P.K. et al., 2004. Emerging infectious diseases of plants: pathogen pollution, climate change and agrotechnology drivers. *Trends in Ecology & Evolution*, **19**(10), pp.535–544.

Anon, 2012. *Phytozome v7.0*. Available at: http://www.phytozome.net/.

Arp, R. & Smith, B., 2008. Function, role, and disposition in Basic Formal Ontology. *Nature Precedings*. Available at: http://hdl.handle.net /10101/ npre.2008.1941.1.

Courtot, M. et al., 2009. MIREOT: the Minimum Information to Reference an External Ontology Term. *Nature Precedings*, (713). Available at: http://precedings.nature.com/documents/3574/version/1.

Cowell, L. & Smith, B., 2010. Infectious Disease Ontology. In *Infectious Disease Informatics*. New York, NY: Springer, pp. 373–395.

Degtyarenko et al. 2007. ChEBI: a database and ontology for chemical entities of biological interest. *Nuc Acids Research* **36**, pp.D344-D350.

FAO, 2012. FAOSTAT, Food and Agricultural Organization of the United Nations. *FAOSTAT*. Available at: http://faostat.fao.org/default.aspx .

Furbank, R.T. & Tester, M., 2011. Phenomics – technologies to relieve the phenotyping bottleneck. *Trends in Plant Science*, **16**(12), pp.635–644.

The Gene Ontology Consortium, 2010. The Gene Ontology in 2010: extensions and refinements. *Nuc Acids Research* **38**, pp.D331-D335.

Gene Ontology, 2012. GO Annotation File Format Guide. Available at: http://www.geneontology.org/GO.format.annotation.shtml.

Giglio, M.G. et al., 2009. Applying the Gene Ontology in microbial annotation. *Trends in Microbiology*, **17**(7), pp.262–268.

Goudie, A. & Cuff, D.J., 2001. *Encyclopedia of global change: environmental change and human society*, Oxford University Press.

Haendel, M. et al., 2007. CARO - The Common Anatomy Reference Ontology. In *Anatomy Ontologies for Bioinformatics: Principles and Practice*. Computational Biology Series. Springer.

Ilic, K., et al. 2007. The Plant Structure Ontology, a Unified Vocabulary of Anatomy and Morphology of a Flowering Plant. *Plant Physiology*: **143**(2), pp.587-599

Jaiswal, P., 2011. Gramene Database: A Hub for Comparative Plant Genomics. In A. Pereira, ed. *Plant Reverse Genetics*. Totowa, NJ: Humana Press, pp. 247–275.

Lin, Y., Xiang, Z. & He, Y., 2011. Brucellosis Ontology (IDOBRU) as an extension of the Infectious Disease Ontology. *Journal of Biomedical Semantics*, **2**(1), p.9.

Manners, J.G., 1993. *Principles of Plant Pathology*, Cambridge, U.K.: Cambridge University Press.

Meehan, T. et al., 2011. Logical development of the Cell Ontology. *BMC Bioinformatics*, **12**(1), p.6.

Mungall, C. et al., 2010. Integrating phenotype ontologies across multiple species. *Genome Biology*, **11**(1), p.R2.

Savage, A., 2000. Changes in MeSH Data Structure, NLM Technical Bulletin. Available at: http://www.nlm.nih.gov/pubs/techbull/ma00/ma00_mesh.html.

Scheuermann, R.H., Ceusters, W. & Smith, B., 2009. Toward an Ontological Treatment of Disease and Diagnosis. *Summit on Translational Bioinformatics*, 2009, p.116.

Smith, B. et al., 2005. Relations in biomedical ontologies. *Genome Biology*, **6**(5), p.R46.

Smith, B. et al., 2007. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nat Biotech*, **25**(11), pp.1251–1255.

Studholme, D.J., Glover, R.H. & Boonham, N., 2011. Application of High-Throughput DNA Sequencing in Phytopathology. *Ann Rev of Phytopathology*, **49**(1), pp.87–105.

Swarbreck, D. et al., 2008. The Arabidopsis Information Resource (TAIR): gene structure and function annotation. *Nuc Acids Research*, **36**(suppl 1), p.D1009 –D1014.

Topalis, P. et al., 2010. IDOMAL: an ontology for malaria. *Malaria Journal*, **9**(1), p.230.

The UniProt Consortium, 2010. Ongoing and future developments at the Universal Protein Resource. *Nuc Acids Research*, **39**(Database), p.D214–D219.

Youens-Clark, K. et al., 2011. Gramene database in 2010: updates and extensions. *Nuc Acids Research*, **39**(suppl 1), p.D1085 –D1094.