Scientific Symposium FAIR Data Sciences for Green Life Sciences
12 December 2018 | Wageningen | the Netherlands

Conference Proceedings | DOI: https://doi.org/10.18174/FAIRdata2018.16287

Linked Data Platform for Plant Breeding and Genomics

Kuzniar, A.[1], G. Singh[2], C. Martinez-Ortiz[1], R.G.F. Visser[2], R. Finkers[2]

[1] Netherlands eScience Center, Science Park 140, 1098XG Amsterdam, the Netherlands
[2] Wageningen University & Research, Droevendaalsesteeg 1, 6708PB Wageningen, the Netherlands
Corresponding author's e-mail: a.kuzniar@esciencecenter.nl

Genetics research is focusing more and more on mining fully sequenced genomes and their annotations to identify genes associated with specific traits (phenotypes) of interest. However, a complex trait is typically associated with multiple quantitative trait loci (QTLs), each encompassing hundreds of genes that positively or negatively affect the desired trait(s). Moreover, the results of QTL mapping studies are commonly described in tables of plant science articles, rather than made accessible in machine-readable formats, which hampers further re-use of these valuable data assets. In the joint NLeSC/WUR project, *candYgene*, we aim to address these needs by developing an analytics platform that makes the integrated geno-/pheno-typic data accessible to plant breeders and researchers interested in *Solanaceae* species [1].

This platform integrates (semi-)structured data from scientific literature and from public molecular biology databases using a Linked Data approach. On the one hand, QTLs were extracted from full-text (open access) articles, as provided by the Europe PMC database, using a recently developed tool called QTLTableMiner++ [2]. Briefly, this tool takes articles in (semi-)structured XML format as input, detects tables with QTL related information, semantically annotates the tables with biological concepts such as trait, gene, protein or genetic marker using domain-specific ontologies (e.g. Crop Ontology, Plant Ontology and Trait Ontology), and outputs the results in machine-readable formats. On the other hand, the genome (proteome) annotations were obtained from the Sol Genomics Network (SGN), UniProt, and Ensemble Plants databases, and were transformed into RDF graphs, including cross-references to other relevant databases (e.g., Gramene, Plant Reactome, InterPro and KEGG). The resulting linked datasets were ingested into the Virtuoso RDF Quad Store and were made available for user queries through a web interface, programmatic SPARQL endpoint or RESTful service.

The linked data platform provides easy access to integrated plant-specific data on genes associated with traits of interest. Our current work focuses on extending the platform with network-based algorithms to enable ranking of candidate genes according to the evidence in multiple data sources. Finally, all research data and associated software developed in the project are being improved and made available according to the FAIR Guiding principles [3].

References

[1] Kuzniar, A. et al. (2018). Linked Data Platform for Plant Breeding & Genomics. *Zenodo*. http://doi.org/10.5281/zenodo.1408128

[2] Singh, G. et al. (2018). QTLTableMiner++: semantic mining of QTL tables in scientific articles. *BMC Bioinformatics*, 19, 183. https://doi.org/10.1186/s12859-018-2165-7

[3] Wilkinson, M. D. et al. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data*, 3, 160018. http://doi.org/10.1038/sdata.2016.18