AgroLD: a Knowledge Graph for the Plant Sciences

Pierre Larmande[1,2,*,†], Bertrand Pitollat[2,3], Ndomassi Tando[1,2], Yann Pomie[1], Bill Happi[1], Valentin Guignon[2,4] and Manuel Ruiz[2,3]

[1]DIADE, IRD, Univ. Montpellier, CIRAD, Montpellier, France
[2]French Institute of Bioinformatics (IFB)—South Green Bioinformatics Platform, Bioversity, CIRAD, INRAE, IRD, Montpellier, France
[3]AGAP, CIRAD, INRAE, Univ. Montpellier, Montpellier, France
[4]Bioversity International, Montpellier, France

Abstract

Recent advances in high-throughput technologies have revolutionized the analysis in the field of the plant sciences. However, there is an urgent need to effectively integrate and assimilate complementary information to understand the biological system in its entirety. We have developed AgroLD, a knowledge graph that exploits Semantic Web technologies to integrate data of interest for the plant science community e.g., rice, wheat, arabidopsis and in this way facilitate the formulation and validation of new scientific hypotheses. AgroLD contains around 900M triples created by annotating and integrating more than 100 datasets coming from 15 data sources. Our objective is to offer a domain specific knowledge platform to answer complex biological and plant sciences questions related to the implication of genes in, for instance, plant disease resistance or adaptative responses to climate change. In this demo, we present some results which currently focused on genomics, genetics and trait associations.

Keywords

Knowledge Graph, Linked Data, FAIR data, Plant Sciences, Bioinformatics

1. The AgroLD Knowledge Graph

AgroLD is built incrementally spanning vast aspects of plant molecular interactions. The current phase covers information on genes, proteins, predictions of homologous genes, metabolic pathways, plant trait associations and genetic studies. At this stage, we have integrated data from several resources such as Ensembl plants, UniProtKB, Gene Ontology Annotation. The choice of these sources has been guided by the biological community, as they are widely used and have a strong impact on the user's confidence. We have also integrated resources developed by the local SouthGreen platform [1] such as TropGeneDB, a tropical plant genetics database, Rice Genome Hub, a rice genomics database, GreenPhylDB, a comparative genomics database for tropical plants, OryzaTagLine, a rice phenotype database and SniPlay, a rice genomic variation database. These resources bring together experimental data produced by researcher groups in

SWAT4HCLS 2023: The 14th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences
*Corresponding author.
pierre.larmande@ird.fr (P. Larmande)
0000-0002-2923-9790 (P. Larmande)

[1]https://www.southgreen.fr

Montpellier and the South of France. The online documentation provides an overview of the integrated data sources [2].
The conceptual framework of AgroLD is based on well-established ontologies in the plant field such as Gene Ontology, Plant Ontology or Plant Trait Ontology. Furthermore, we developed a dedicated schema [3] that creates links between the imported ontologies and introduces new classes and properties. The online documentation shows the complete list of the used ontologies. The majority of these ontologies are hosted by the OBO Foundry project.

2. Statistics

As of today, AgroLD contains more than 900 Millions triples resulting of the integration of roughly 100 datasets gathered in 33 named graphs. Table 1 gives an overview of available resources and tools. All datasets are available in Zenodo under the Creative Commons Attribution 4.0 International license (CC-BY 4.0). Each resource can contain several datasets, for instances, one dataset per species or per data type. Combining all ontologies and datasets imported, AgroLD graph gather 383 classes and 793 properties. Among the pipelines developed to lift up the datasets, we focused also on connecting our datasets with others. The property *rdfs:seeAlso* reach the total number of almost 80 millions of outbound links making the AgroLD graph correctly linked with other datasets in the LOD. Besides, we paid attention to increasing the number of semantic annotations with imported ontologies, which increased the number of links between datasets making the overall graph denser. We created more than 14 million semantic links linking entities to ontological classes. Finally, our data linking strategy allowed us to create around 160,000 *owl:sameAs* links between entities.

Table 1
Links to AgroLD resources and tools

Name of resource or tool and description, URL
Data
AgroLD datasets, https://doi.org/10.5281/zenodo.4694518
List of graphs, http://www.agrold.org/documentation.jsp
List of ontologies, http://www.agrold.org/documentation.jsp
AgroLD vocabulary, https://github.com/SouthGreenPlatform/AgroLD_ETL/tree/master/model
AgroLD SPARQL Endpoint, http://agrold.southgreen.fr/sparql
Example queries, http://www.agrold.org/sparqleditor.jsp
Tools
Web application, https://github.com/SouthGreenPlatform/AgroLD_webapp
RDF conversion pipelines (GFF2RDF, GAF2RDF, VCF2RDF, Datasets), https://github.com/SouthGreenPlatform/AgroLD_ETL
Publications
Original paper, https://doi.org/10.1371/journal.pone.0198270
Resource paper, https://doi.org/10.1007/978-3-030-88361-4_29

[2] http://www.agrold.org/documentation.jsp
[3] https://github.com/SouthGreenPlatform/AgroLD_ETL/tree/master/model