

# Exposing French agronomic resources as Linked Open Data

Aravind Venkatesan<sup>1</sup>, Nordine El Hassouni<sup>2</sup>, Florian Phillippe<sup>3</sup>, Cyril Pommier<sup>3</sup>, Hadi Quesneville<sup>3</sup>, Manuel Ruiz<sup>1,2</sup>, Pierre Larmande<sup>1,4,5</sup>

<sup>1</sup>Institut de biologie Computationelle, Montpellier, France  
Aravind.Venkatesan@lirmm.fr

<sup>2</sup>UMR AGAP, CIRAD, Montpellier, France  
{nordine.el\_hassouni,manuel.ruiz}@cirad.fr

<sup>3</sup>URGI, INRA, Versailles, France  
{fphilippe,Cyril.pommier,hadi.quesneville}@versailles.inra.fr

<sup>4</sup>UMR DIADE, IRD, Montpellier, France

<sup>5</sup>Equipe Zenith, INRIA et LIRMM, Montpellier, France  
pierre.larmande@ird.fr

**Abstract.** The advancements in empirical technologies has generated vast amounts of heterogeneous data. This situation has created a need to integrate the data to understand the system of interest in its entirety. Therefore, information systems play a crucial role in managing these data, enabling the biologists in the extraction of new knowledge. The plant bioinformatics node of the Institut Français de Bioinformatique (IFB) maintains public information systems that houses domain specific data. Currently, efforts are being taken to expose the IFB plant bioinformatics resources as Linked Open Data, utilising domain specific ontologies and metadata. Here, we present the overview and the initial results of the project.

**Keywords:** Data integration, Data interoperability, Knowledge management, Linked Data, RDF, Bioinformatics application, Agronomic research

## 1 Introduction

Agronomy is an overarching field that encompasses various research areas such as genetics, plant molecular biology, and agro-ecology. The last several decades has seen the successful implementation of high-throughput technologies that have revolutionised research in agronomy. These technological advancements have resulted in a number of initiatives been taken to systematically store and share information over the web. Among others, the Institut Français de Bioinformatique (IFB) is a French-ELIXIR node (<http://www.elixir-europe.org/about/elixir-france>) that is focused on providing integrated services for the life science community. The plant research wing of the IFB platform provides access to a number of databases and tools distributed across six regional bioinformatics portals. Data in these portals ranges from 'omics' to

genetics (genetic markers, maps and phenotypes) for various crop species. The objective of the current effort is to develop RDF knowledge bases that integrates existing domain specific ontologies and data from the respective regional portals, promoting data interoperability between the resources. To this end, we have developed the Agronomic Linked Data knowledge base ([www.agrold.org](http://www.agrold.org)) that is representative of the data housed in the southern region portal of France, the SouthGreen Bioinformatics platform (SG) (<http://www.southgreen.fr/>).

## 2 AgroLD for SG resources

Currently, SG consists of 12 databases covering various plant species such as Banana, Cocoa, Maize and Rice. AgroLD is being developed in phases to expose all of these databases as Linked Data. Currently, Phase I of AgroLD includes data from:

1. TropGeneDB (Hamelin et al. 2013), a database that hosts genetic, molecular and phenotypic information on tropical crop species.
2. OryGenesDB (Droc et al. 2006), a database that serves as a repository on functional genomics for rice.
3. Oryza Tag Line (Larmande et al. 2008), a database that contains sequence information (Flanking Sequence Tags) that are based on molecular categorisation of mutagen insertion sites for rice.
4. GreenPhylDB (Conte et al. 2008), provides sequence homology information for the members of kingdom *plantae*.

Additionally, domain specific ontologies, ontology annotations, proteomics and genomics information from a variety of publically available data sources have been integrated, this includes Gene Ontology, Plant Ontology, UniprotKB, Gramene (Gene, ontology annotation, gene, Quantitative Trait Loci (QTL) and Metabolic Pathway information) (Monaco et al. 2014). The objective of this is to provide the critical mass required to implement real world use cases. Currently, AgroLD includes data pertaining to selected species namely, *Oryza* species (*O.sativa*, *O.barthii*, *O.brachyantha*, *O. glaberrima* and *O.meridionalis*), *Arabidopsis thaliana*, *Sorghum bicolor*, *Zea mays* and *Triticum* species (*T.aestivum* and *T. uraruta*). In the subsequent phases information pertaining to other species and SG databases will be considered. The AgroLD effort will be further extended set-up RDF knowledge bases to host data from other regional portals.

## References

- Barrell, D. et al., 2009. The GOA database in 2009 - An integrated Gene Ontology Annotation resource. *Nucleic Acids Research*, 37(SUPPL. 1).
- Conte, M.G. et al., 2008. GreenPhylDB: A database for plant comparative genomics. *Nucleic Acids Research*, 36(SUPPL. 1).
- Droc, G. et al., 2006. OryGenesDB: a database for rice reverse genetics. *Nucleic acids research*, 34(Database issue), pp.D736–D740.
- Hamelin, C. et al., 2013. TropGeneDB, the multi-tropical crop information system updated and extended. *Nucleic Acids Research*, 41(D1).
- Larmande, P. et al., 2008. Oryza Tag Line, a phenotypic mutant database for the Génoplatte rice insertion line library. *Nucleic Acids Research*, 36(SUPPL. 1).
- Monaco, M.K. et al., 2014. Gramene 2013: Comparative plant genomics resources. *Nucleic Acids Research*, 42(D1).