Exposing French agronomic resources as Linked Open Data

Aravind Venkatesan[1], Nordine El Hassouni[2], Florian Phillipe[3], Cyril Pommier[3], Hadi Quesneville[3], Manuel Ruiz[12], Pierre Larmande[145]

[1] Institut de biologie Computationelle, Montpellier, France
Aravind.Venkatesan@lirmm.fr
[2] UMR AGAP, CIRAD, Montpellier, France
{nordine.el_hassouni,manuel.ruiz}@cirad.fr
[3] URGI, INRA, Versailles, France
{fphilippe,Cyril.pommier,hadi.quesneville}@versailles.inra.fr
[4] UMR DIADE, IRD, Montpellier, France
[5] 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

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 such as Gramene (Monaco et al., 2014), TAIR (Lamesch et al., 2012), OryzaBase (Kurata et al., 2006), Plant Reactome (Croft et al., 2014), GnpIS (Steinbach et al., 2013) and the South Green bioinformatics platform (http://www.southgreen.fr), to name a few.

However, using these resources comprehensively, taking advantage of the associated cross-disciplinary research opportunities poses a major challenge to both domain scientists and information technologists. Effective data integration and management allows a broader perspective across many disciplines, than is possible from one or a series of individual studies. A solution for the data integration challenges is offered by the Semantic Web (SW) technologies (Berners-Lee & Hendler 2001). 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) that is representative of the data housed in the southern region portal of France, the SouthGreen Bioinformatics platform (SG) (http://www.southgreen.fr/).

Semantification of the IFB plant bioinformatics nodes

Institut Français de Bioinformatique (IFB) is a French national node (http://www.elixir-europe.org/about/elixir-france) that is focused on providing integrated services for the life science community. The IFB platform provides access to databases, tools and services that covers three main domains namely, microbial, plant and health sciences. The IFB IT infrastructure is linked to six regional bioinformatics centers, the ReNaBi (French Bioinformatics Platforms Network), representing various regions of the French territory (ReNaBi-NE, North-East; PRABI, Rhône-Alpes region; ReNaBi-GS, Great South; ReNaBi-SO, South-West; ReNaBi-GO, Great West and APLIBIO, Paris area). These six regional centers are consists of regional bioinformatics platforms (PFs). Taken together, IFB will represent France in the ELIXIR European infrastructure initiative. To this end, the plant bioinformatics PFs maintain public data repositories that ranges from 'omics' to genetic data (genetic markers, maps and phenotypes) for various crop species.

Currently, the plant-centric PFs are working towards exposing their resources as linked data. The objective of the current effort is to develop RDF knowledge base that integrates existing domain specific ontologies and data from the respective PFs. This will promote interoperability between the databases. In the initial phase, two representative PFs are involved in this semantification process, namely:

a) The *Unité de Recherche Génomique-Info* (URGI) platform (https://urgi.versailles.inra.fr/) associated with the *Institut National de la Recherche Agronomique* (INRA), dedicated to maintain curated information on plants and crop parasite. The platform is part of the APLIBIO ReNaBi and plays a key role in the Wheat Initiative (http://wheatis.org/).

b) The South Green Bioinformatics platform (SG) part of the ReNaBi GS mainly associated with *Centre de coopération internationale en recherche agronomique pour le développement* (CIRAD) and *Institut de recherche pour*

le développement (IRD) among other regional institutes. SG provides tools and databases dedicated for genomic resource analysis of southern and Mediterranean plants.

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. glaberimma* 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

Berners-Lee T. & Hendler J. 2001. Publishing on the semantic web. Nature, 410, 1023-4.

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).

Croft D. et al, (2014). The Reactome pathway knowledgebase. Nucleic Acids Research. 42(D1), D472-D477.

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).

Kurata N. & Yamazaki Y. (2006). Oryzabase. An integrated biological and genome information database for rice. Plant physiology. 140(1), 12-17.

Lamesh P. et al. (2012). The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. Nucleic acids research. 40(D1), D1202-D1210.

Larmande, P. et al., 2008. Oryza Tag Line, a phenotypic mutant database for the Génoplante 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).

Steinbach D., et al., 2013. GnpIS: an information system to integrate genetic and genomic data from plants and fungi. Database, 2013. bat058.