Exposing French agronomic resources as Linked Open Data

Aravind Venkatesan, Nordine El Hassouni, Florian Phillipe, Cyril Pommier, Hadi Quesneville, Manuel Ruiz, Pierre Larmande

To cite this version:

Aravind Venkatesan, Nordine El Hassouni, Florian Phillipe, Cyril Pommier, Hadi Quesneville, et al.. Exposing French agronomic resources as Linked Open Data. IN-OLIVE, Jun 2016, Montpellier, France. hal-01411759

HAL Id: hal-01411759
https://hal.archives-ouvertes.fr/hal-01411759
Submitted on 7 Dec 2016

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L’archive ouverte pluridisciplinaire HAL, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d’enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.
Exposing French agronomic resources as Linked Open Data

Aravind Venkatesan1, Nordine El Hassouni2, Florian Philippe3, Cyril Pommier3, Hadi Quesneville3, Manuel Ruiz12, Pierre Larmande145

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. urarata). 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**


