Towards efficient data integration and knowledge management in the Agronomic domain

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

Résumé : Today, the revolution in empirical technologies has generated vast amounts of data. This data deluge has created an urgent need to assimilate it with a panoramic view. To this end, information systems play a central role in managing and integrating these data, aiding the biologists in exploiting this integrated information for the extraction of new knowledge. The plant bioinformatics node of the Institut Français de Bioinformatique (IFB) maintains public information systems where a variety of domain specific data are integrated. Currently, efforts are being taken to expose the IFB plant bioinformatics resources as RDF, utilising domain specific ontologies and metadata. Here, we present the overview and the progress of the project.

Mots-clés: Data integration, data interoperability, knowledge management, Semantic Web, RDF, Bioinformatic application, Agronomic research

1 Introduction

Agronomy is an overarching field constituting various research areas such as genetics, plant molecular biology, ecology and earth science. The last several decades has seen the successful development of high-throughput technologies that have revolutionised and transformed agronomic research. The application of these technologies have generated large quantities of data. 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.

The definitive aim of agronomic research being the improvement of crop production through sustainable methods. It is important to efficiently over lay research findings from the allied fields, by automating data analyses for hypotheses generation, ultimately reducing the knowledge discovery cycle. 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. In the long run, this allows information to be used for purposes other than those for which they were originally intended, to address questions that were unapproachable at the time the data were collected. To this end, the need for an umbrella approach for providing uniform data is a much discussed topic. For instance, the Research Data Alliance (RDA, https://rd-alliance.org/) through its interest group namely, the Agriculture Data Interoperability Interest group (https://rd-alliance.org/groups/agriculture-data-interest-group-igad), has created a platform to discuss the need to improve data exchange enabling data integration in this domain.

A solution for the data integration challenges is offered by the Semantic Web (SW) technologies (Berners-Lee & Hendler 2001). SW was proposed, to remedy the fragmentation of all the potentially useful information dispersed over the web. This is founded on a stack of technologies such as the Resource Description Framework (RDF, www.w3.org/RDF/), RDF Schema (RDFS, http://www.w3.org/TR/rdf-schema/), Web Ontology Language (OWL, www.w3.org/TR/owl-features/) and the SPARQL Query Language (SPARQL, www.w3.org/TR/rdf-sparql-query/). With ontologies providing the knowledge scaffold, a successfully implemented SW application can be used by scientists to pose complex questions that would then assess and return highly relevant answers to those questions. This is useful, as the assessment of biological findings against prior knowledge, after which the best supported hypotheses can be selected for further testing.

We are currently witnessing a growing acceptance of SW (RDF in particular) for the management of disparate biological databases in the bio-medical field. Several projects have been undertaken to demonstrate the potential of SW, some notable initiatives include Bio2RDF (Belleau et al., 2008), BioGateway (Antezana et al., 2009), Linked Life Data (Momtchev et al., 2009), KUPKB (Jupp et al., 2011) and OpenPHACTS (Williams et al., 2012). These initiatives have demonstrated the advantages of SW including rich knowledge representation, streamlined data integration and optimised querying. Moreover, primary data providers such as EMBL-EBI (https://www.ebi.ac.uk/rdf/platform), UniprotKB (http://beta.sparql.uniprot.org/) and NCBI (Anguita et al., 2013) are also making their data available as RDF.

Presently, in the agronomic domain plant centric ontologies such as Plant Ontology (PO), Plant Trait Ontology (TO) and Plant Environment Ontology (EO) (http://planteome.org/), are being used by various databases as a method to provide cross-domain common entry points. Nevertheless, efficient knowledge management additionally requires information to be represented in a machine-readable form. Unlike the bio-medical domain, the agronomic sciences is yet to exploit the full potential of SW. Therefore, initiatives to build on previous efforts to expose agronomy data on the SW is essential. Furthermore, attempts have to be made to pursue various joint collaborations with the intended stakeholders (for instance, plant biologists and breeders) to bridge the gap between SW and the opportunities that come along with it (Venkatesan et al. 2014).

2 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 (see Figure 1). 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.

Additionally, future efforts will be taken provide RESTful APIs to make the RDF repositories accessible from within workflow environment such as Taverna (Wolstencroft et al., 2013) and Galaxy (Giardine et al., 2005).

3 RDF store integrating South Green resources

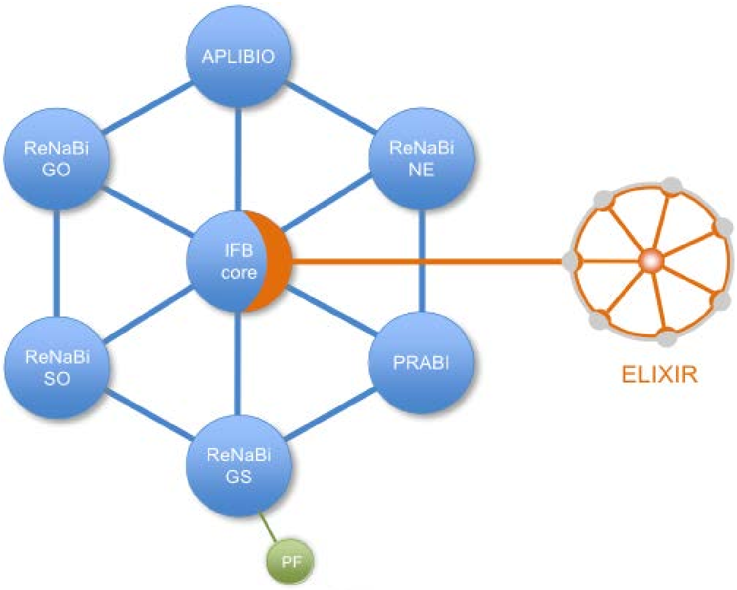
This section provides description of the effort taken to establish the RDF repositories representing the IFB plant PFs. As a primary step, resources hosted under the South Green bioinformatics platform was chosen for modeling data as RDF.

3.1 Design principles

Currently, SG houses 12 databases covering various plant species such as Banana, Cocoa, Maize and Rice. The RDF store is designed to provide plant biologists with a knowledge system that captures knowledge represented in these databases. The following design principles were followed in the RDF store development:

1. Developing the store in various phases.
2. Integrating domain specific data and ontologies to enable comparative analyses.
3. Providing maximum flexibility both for end users and for future extensions.
4. Making the store Linked Data compliant.

Figure 1 - The illustration shows the structure of IFB node with ELIXIR. The blue nodes represent regional platforms (ReNaBi) across France. The green node represent the regional bioinformatics platform of ReNaBi (Figure adapted from Perriere 2012.).



3.2 Data sources

The phase I of the RDF store development was focused on integrating a number of well-established plant centric ontologies, this includes the Gene Ontology (three variants) (Ashburner et al., 2000), PO, TO and EO. For phase I, the SG databases that were included are:

1. TropGeneDB (Hamelin et al., 2012), 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. GreenPhyl (Conte et al., 2008), provides sequence homology information for the members of kingdom *plantae*.

Furthermore, ontology annotations, proteomics and genomics information from a variety of publically available data sources were integrated, this includes, UniprotKB (Magrane & Uniprot Consortium 2011), GOA (Barrell et al., 2009), Gramene (ontology annotation, gene and Quantitative Trait Loci (QTL) information), AraCyc (Mueller et al., 2003), RiceCyc, SorghumCyc, and MaizeCyc (Dharmawardhana et al., 2013) and OryzaBase. The integration of additional data resources provides the critical mass required for implementing real world use cases. Currently, the RDF store is limited to a 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*). Table 1 provides a detailed list of data resources integrated in SG RDF store.

TABLE 1 – The table shows the breakdown of the data sources integrated into SG RDF store with the corresponding species.

Species **Data Sources**		**Oryza spp.**	**A.thaliana**	**S.bicolor**	**Z.mays**	**Triticum spp**
SG platform	TropGeneDB	✓			✓	✓
	OryGenesDB	✓	✓	✓		
	Oryza Tag Line	✓				
	GreenPhyl	✓	✓	✓	✓	
Ontology associations	GOA	✓	✓	✓	✓	✓
	Gramene-PO	✓	✓	✓	✓	✓
	Gramene-TO	✓				
	Gramene-EO	✓				
Gramene genes		✓	✓	✓	✓	✓
Gramene QTL		✓				
UniProtKB		✓	✓	✓	✓	✓
AraCyc			✓			
RiceCyc		✓				
SorghumCyc				✓		
MaizeCyc					✓	

3.3 South Green RDF store construction

The SG RDF store is built using a semi-automated pipeline implemented in Python ver.2.7. The pipeline integrates information from SG resources as well as other resources (refer section: 3.2 Data sources) as RDF. The conversion of these resources was performed with newly developed parsers, for GOA, BioPython (www.biopython.org/) module (Bio.Uniprot.GOA) was used. To avoid duplication, information that were available as RDF was directly utilized such as, the candidate ontologies. The RDF graphs have been loaded into Open Link Virtuoso (http://virtuoso.openlinksw.com). The SPARQL endpoint is currently being tested and can be accessed at: http://volvestre.cirad.fr:8890/sparql.

In order to make SG RDF store Linked Data complaint, dereferenceable stable URIs provided by Identifiers.org (http://identifiers.org/) and Ontobee (www.ontobee.org/) were used. The conversion new datasets that are not included in these registries, new URIs were minted with a common name-space: *http://www.southgreen.fr/agrold/*. The identifiers for these datasets take the form *http://www.southgreen.fr/agrold/**[resource_namespace]/[identifier]***, for example, AraCyc describes metabolic pathway information associated with *A.thaliana* genes. Thus, the URI for AraCyc pathway identifier would be: *http://www.southgreen.fr/agrold/aracyc.pathway/PWYQT-4482*. Similarly, for resources that requiring new properties, would be of the form http://www.southgreen.fr/agrold/**[vocabulary]/[property]**. For example, Gramene provides protein-EO annotations, the property *expressed_in* is used link the two resources. The absolute URI for this property would be: *http://www.southgreen.fr/agrold/vocabulary/expressed_in*.

3.4 Querying

In this section we demonstrate the utility of the knowledge base with the help of a few example SPARQL queries. These queries will be made available as a part of a list of sample queries provided on dedicated query page (under construction).

Q1. Retrieve the local neighbourhood of Oriza sativa japonica protein: IAA16 - Auxin-responsive protein (UniProt accession: P0C127).

SPARQL query:

```
BASE <http://www.southgreen.fr/agrold/>
PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
PREFIX obo:<http://purl.obolibrary.org/obo/>
PREFIX uniprot:<http://purl.uniprot.org/uniprot/>
PREFIX vocab:<vocabulary/>
PREFIX graph:<protein.ontology.associations>

SELECT distinct ?predicate ?object ?object_label
WHERE {
 GRAPH graph: {
  uniprot:P0C127 ?predicate ?object.
  OPTIONAL {
   GRAPH ?g {
    ?object rdfs:label ?object_label.}}}}
```

Q2. Retrieve genes that participate in a pathway: Calvin cycle.

SPARQL query:

```
BASE <http://www.southgreen.fr/agrold/>
PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
PREFIX obo:<http://purl.obolibrary.org/obo/>
PREFIX uniprot:<http://purl.uniprot.org/uniprot/>
PREFIX vocab:<vocabulary/>
PREFIX graph:<ricecyc>
PREFIX   pathway:<http://www.southgreen.fr/agrold/ricecyc.pathway/CALVIN-
PWY>

SELECT DISTINCT ?gene ?name
WHERE {
 GRAPH graph: {
  pathway: vocab:has_agent ?gene.
  ?gene rdfs:label ?name.}}
```

These queries offer a glimpse of the range of biological questions that can be addressed to the knowledge base. Since the data is represented as linked data users could also query the knowledge base in combination with complimentary RDF knowledge bases using query federation features offered by SPARQL ver.1.1.

4 Future Directions

SG RDF store currently hosts information for select species (refer section: 3.2 Data sources), in the subsequent phases information pertaining to other species such as Banana and Cocoa will be integrated. Additionally, the integration of other databases under SG that includes SNiPlayDB (Dereeper et al., 2011), CocoaGenDB (Argout et al., 2007) and EuriGen (Courtois 2012) will be considered. Presently, semantic web knowledge bases are accessed via a SPARQL endpoint. However, these endpoints are more machine-friendly and users are required at the minimum a moderate knowledge of SPARQL to exploit the integrated information. Obviously, this is not an optimal solution for users not acquainted with SPARQL. Hence efforts will be made to develop user friendly query interface that is optimized to aid non-technical users. Furthermore, collaborations with domain experts will be pursued to develop real world use cases to demonstrate the advantages of SW. To this end, the RDF store will be augmented with additional resources to suit the use cases. Furthermore, this effort will be extended to the information hosted at URGI.

5 Conclusion

The drastic increase in the amount of data generated in the agronomic domain requires efficient knowledge management practices. SW certainly provides a robust method to integrate information representing domain specific knowledge. Exposing the IFB plant PFs as RDF will aid both domain experts and bioinformaticians to take advantage of the integrated information. With the development of SG RDF store, we have taken the initial steps towards this goal.

Références

Anguita, A., Garcia-Remesal, M., de la Iglesia, D., & Maojo, V. (2013). NCBI2RDF: enabling full RDF-based access to NCBI databases. BioMed Research International. 983805.

Antezana E., Blondé W., Egaña M., Rutherford A., Stevens, R. De Baets B., Mironov V. & Kuiper M. (2009). BioGateway: a semantic systems biology tool for the life sciences. BMC bioinformatics, 10(Suppl 10), S11. Bouaud J. (1978a). De la façon d'écrire un article. Le Journal LaTeX. 2, p. 101-105.

Application Delivery Strategies published by META Group Inc.

Argout X. Ruiz M. Rouard M. Turnbull C. Lanaud C. Rosenquist E. et al. CocoaGen DB: a Web portal for crossing cocoa phenotypic, genetic and genomic data from ICGD and TropGeneDB databases.CocoaGen DB: a Web portal for crossing cocoa phenotypic, genetic and genomic data from ICGD and TropGeneDB databases. In: 15th International Cocoa Research Conference. Vol 1. San Jose, Costa Rica; 2007. p. 515-8

Ashburner et al. Gene ontology: tool for the unification of biology (2000) *Nat Genet* **25(1)**:25-9.

Barrell, D., Dimmer, E., Huntley, R. P., Binns, D., O'Donovan, C., & Apweiler, R. (2009). The GOA database in 2009—an integrated Gene Ontology Annotation resource. Nucleic acids research, 37(suppl 1), D396-D403.

Belleau F., Nolin M. A., Tourigny N., Rigault P., & Morissette J. (2008). Bio2RDF: towards a mashup to build bioinformatics knowledge systems. Journal of biomedical informatics, 41(5), 706-716.

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

Conte, M. G., Gaillard, S., Lanau, N., Rouard, M., & Perin, C. (2008). GreenPhylDB: a database for plant comparative genomics. Nucleic acids research, 36(suppl 1), D991-D998.

Courtois, B., Frouin, J., Greco, R., Bruschi, G., Droc, G., Hamelin, C. & Ahmadi, N. (2012). Genetic diversity and population structure in a European collection of rice. Crop science, 52(4), 1663-1675.

CROFT D., MUNDO A. F., HAW R., MILACIC M., WEISER J., WU G., CAUDY M., GARAPATI P., GILLESPIE M., KAMDAR M R., JASSAL B., JUPE S., MATTHEWS L., MAY B., PALATNIK S., ROTHFELS K., SHAMOYSKY V., SONG H., WILLIAMS M., BIRNEY E., HERMJAKOB H., STEIN L. & D'EUSTACHIO P. (2014). The Reactome pathway knowledgebase. Nucleic Acids Research. 42(D1), D472-D477.

DEREEPER, A., NICOLAS, S., LE CUNFF, L., BACILIERI, R., DOLIGEZ, A., PEROS, J. P. & THIS, P. (2011). SNiPlay: a web-based tool for detection, management and analysis of SNPs. Application to grapevine diversity projects. BMC bioinformatics, 12(1), 134.

DHARMAWARDHANA, P., REN, L., AMARASINGHE, V., MONACO, M., THOMASON, J., RAVENSCROFT, D. & JAISWAL, P. (2013). A genome scale metabolic network for rice and accompanying analysis of tryptophan, auxin and serotonin biosynthesis regulation under biotic stress. Rice, 6(1), 15.

DROC, G., RUIZ, M., LARMANDE, P., PEREIRA, A., PIFFANELLI, P., MOREL, J. B. & PERIN, C. (2006). OryGenesDB: a database for rice reverse genetics. Nucleic acids research, 34(suppl 1), D736-D740.

GIARDINE, B., RIEMER, C., HARDISON, R. C., BURHANS, R., ELNITSKI, L., SHAH, P. & NEKRUTENKO, A. (2005). Galaxy: a platform for interactive large-scale genome analysis. Genome research, 15(10), 1451-1455.

HAMELIN, C., SEMPERE, G., JOUFFE, V., & RUIZ, M. (2012). TropGeneDB, the multi-tropical crop information system updated and extended. Nucleic acids research, gks1105.

JUPP, S., KLEIN, J., SCHANSTRA, J., & STEVENS, R. (2011). Developing a kidney and urinary pathway knowledge base. J. Biomedical Semantics, 2(S-2), S7.

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

LAMESCH P., BERARDINI T Z., LI D., SWARBRECK D., WILKS C., SASIDHARAN R. & HUALA E. (2012). The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools. Nucleic acids research, 40(D1), D1202-D1210.

LARMANDE, P., GAY, C., LORIEUX, M., PÉRIN, C., BOUNIOL, M., DROC, G. & GUIDERDONI, E. (2008). Oryza Tag Line, a phenotypic mutant database for the Genoplante rice insertion line library. Nucleic acids research, 36(suppl 1), D1022-D1027.

MAGRANE, M., & UNIPROT CONSORTIUM. (2011). UniProt Knowledgebase: a hub of integrated protein data. Database, 2011, bar009.

MOMTCHEV, V., PEYCHEV, D., PRIMOV, T., & GEORGIEV, G. (2009). Expanding the pathway and interaction knowledge in linked life data. Proc. of International Semantic Web Challenge.

MONACO K M., STEIN J., NAITHANI S., Wei, DHARMAWARDHANA P., KUMARI S., AMARASINGHE V., CLARK K J., THOMASON J., PREECE J., PASTERNAK S., OLSON A., JIAO Y. LU Z., BOLSER D., KKERHORNOU A., STAINES D., WALTS B., WU G., D'EUSTACHIO P., HAW R., CROFT D., KERSEY J P., STEIN L., JAISWAL P. &WARE D. (2014). Gramene 2013: comparative plant genomics resources. Nucleic Acids Reseach. 42 (D1): D1193-D1199.

MUELLER, L. A., ZHANG, P., & RHEE, S. Y. (2003). AraCyc: a biochemical pathway database for Arabidopsis. Plant Physiology, 132(2), 453-460.

PERRIERE, G. (2012). ReNaBi-IFB: The French Bioinformatics Infrastructure. EMBnet journal, North America, 18, jun. 2012

STEINBACH D., ALAUX M., AMSELEM J., CHOISNE N., DURAND S., FLORES R. KELIET A., KIMMEL E., LAPALU N., LUYTEN I., MICHOTEY C., MOHELLIBI N., POMMIER C., REBOUX S., VALDENAIRE D., VERDELET D. & QUESNEVILLE, H. (2013). GnpIS: an information system to integrate genetic and genomic data from plants and fungi. Database, 2013. bat058.

VENKATESAN, A., TRIPATHI, S., DE GALDEANO, A. S., BLONDÉ, W., LÆGREID, A., MIRONOV, V., & KUIPER, M. (2014). Finding gene regulatory network candidates using the gene expression knowledge base. BMC bioinformatics, 15(1), 386.

WILLIAMS, A. J., HARLAND, L., GROTH, P., PETTIFER, S., CHICHESTER, C., WILLIGHAGEN, E. L. & MONS, B. (2012). Open PHACTS: semantic interoperability for drug discovery. *Drug discovery today*, *17*(21), 1188-1198.

WOLSTENCROFT, K., HAINES, R., FELLOWS, D., WILLIAMS, A., WITHERS, D., OWEN, S. & GOBLE, C. (2013). The Taverna workflow suite: designing and executing workflows of Web Services on the desktop, web or in the cloud. Nucleic acids research, gkt328.