







nstitut de recherche pour le développement

Agronomic Linked Data (AgroLD)

a knowledge system to enable integrative biology in Agronomy

Aravind Venkatesan¹, Nordine El Hassouni², Gildas Tagny¹, Manuel Ruiz^{1,2} and Pierre Larmande^{1,3}

¹Institut de biologie Computationnelle, Montpellier; ²UMR AGAP, CIRAD, Montpellier; ³UMR DIADE, IRD, Montpellier, France

Background

The advances in empirical technologies have 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. We have developed the Agronomic Linked Data knowledge base, a knowledge system that exploits the Semantic Web technology to integrate information on plant species widely studied by the agronomic research community. The objective of this effort is to provide the community with a platform for domain specific knowledge, capable of answering complex biological questions and in this way facilitating the formulation of new hypotheses. Here we present the initial results of the Agronomic Linked Data project, phase one of the project being focused on integrating genomics, proteomics and phenomics information.



The table shows all data and ontologies integrated in the triple Store



AgroLD hosts South Green (www.southgreen.fr) data, along with other publically available data sources (e.g: Gramene).

Sources	URL s	# tuples	Crops	Ontologies used	# triples produced
Ontology associations	geneontology.org	1 160K	R, W, A, M, S	GO, PO, TO, EO	2 700K
Gramene	gramene.org	1 718K	R, W, M, A, S	GO, PO, TO, EO	5 172K
UniprotKB	uniprot.org	1 400K	R, W, A, M, S	GO, PO	10 000 K
OryGenesDB	orygenesdb.cirad.fr	1 100K	R, S, A,	GO, SO	2 300K
Oryza Tag Line	oryzatagline.cirad.fr	22K	R	PO, TO	300K
TropGeneDB	tropgenedb.cirad.fr	2k	R	PO, TO	20K
GreenPhylDB	greenphyl.org	100K	R, A	GO, PO	700K
SniPlay	sniplay.southgreen.fr	2 000K	R	GO	16 000K





R=rice, W=wheat, A=Arabidopsis, S= sorghum, M= maize

GO = Gene Ontology, PO = Plant Ontology, TO = Plant Trait Ontology, EO = Environment Ontology, SO = Sequence Ontology

Species specific break down of the data sources

AgroLD schema and ETL



understanding the underlying knowledge.

The *Explore Relationships* tool aids in exploring relationships between existing entities.

Agronomic Linked Data (AgroLD)

Home Search - Help - About Please send us your feedback! 🔶

The Agronomic Linked Data (AgroLD) Project

Query Patterns

i. Retrieve list of graphs (selec

activity" (select)

Search terms by label (select)

List relation types in a given graph (select)

Get the ID corresponding to the ontology term "ho

0. Get the level 4 ancestor of GO:0004409 (select)

Get the level 2 descendence of GO:0003824 (

Get protein ids associated with the ontological id G

. Retrieve the local neighbourhood of Oryza sativa japonica protein: IAA16

Identify Wheat proteins that are involved in root development. (select

Retrieve genes that participate in a given pathway: Calvin cycle (selec Retrieve Proteins associated with a given QTL: DTHD (days to heading

Get the name of the ontological element that has the ID "GO:000382

uxin-responsive protein (UniProt accession:P0C127) (select)

At the Institute of Computational Biology (IBC), we are involved in developing methods to aid data integration and knowledge management within the plant biology domain to improve information accessibility of heterogeneous data. Among others, a solution for the data integration challenges is offered by the Semantic Web technologies. The semantic web has emerged as one of the most promising solutions for high scale integration of distributed resources. This is made possible by a stack of technologies such as the Resource Description Framework (RDF), RDF Schema (RDFS), Web Ontology Language (OWL) and the SPARQL Query Language (SPARQL) proposed by the World Wide Web Consortium (W3C). RDF forms the basis of the stack allows modeling information as a directed graph composed of triples that can be queried using SPARQL.

AgroLD is a RDF knowledge base that consists of data integrated from a variety of plant resources and ontologies. The aim of the Agronomic Linked Data (AgroLD) project is to provide a portal for bioinformatics and domain experts to exploit the homogenized data models towards efficiently generating research hypotheses.

Quick Search	Advanced Search	Explore Relationships	SPARQL Query Editor
Search with keywords and browse AgroLD Knowledge Base	Search with keywords, browse, and get answers to some biological questions	Search easily existing relationships between entities	Edit and submit your SPARQL Queries to the sparql endpoint of AgroLD located at <u>http://volvestre.cirad.fr:8890/sparql</u>



Agronomic Linked Data (AgroLD)





Execution timeout 20000 mi	illiseconds (values less than 1000 are igr	lored) Results Format RDF/XML • Download Result	13. Get QTL Ids associated with 14. Describe uniprot: P0C127 (s	the ontological id EO:0007403 (select)
Filename to Save As: query.spar	rql Save Query	Choose File No file cho	File	Parta-ca)
Results				
Raw Response Table	Pivot Table	•	Search	Show 50 v entries
gene		∂ name	d taxon name	
1 http://identifiers.org/ensem	nbl.plant/AT1G18270	fructose-bisphosphate aldolase	obo:NCBITaxon_3702	
2 http://identifiers.org/ensem	nbl.plant/AT1G42970	glyceraldehyde-3-phosphate dehydrogenase	obo:NCBITaxon_3702	
3 http://identifiers.org/ensem	nbl.plant/AT1G43670	fructose-1,6-bisphosphatase	obo:NCBITaxon_3702	
el EnsemblPlants	S ▼ BLAST BioMart <u>Tools D</u>	ownloads Documentation Website help		Login/Register
Arabidopsis thaliana (TAIR10) 🔻	Location: 1:6,283,412-6,293,871 Get	ne: AT1G18270		
Sene-based displays - Summary	Gene: AT1G18270			
 Splice variants Transcript comparison 	Description	ketose-bisphosphate aldolase class-II family protein [Source:TAIR;Acc:AT1G18270	el]	
 Supporting evidence Gene alleles 	Location	Chromosome 1: 6,283,412-6,293,871 reverse strand.		
3 Sequence	About this gene	This gene has 3 transcripts (splice variants), 37 orthologues and 6 paralogues.		
- Gene families	Transcripts	Show transcript table		
 External references Regulation Literature 	Summary 🕑			
Ontology GO: biological process	Gene type	Protein coding		
 GO: molecular function PO: plant structure developmer 	Annotation Method	Gene annotation by $\underline{TAIR}_{\mathbf{B}}$ through a process of automatic and manual curation.		
PO: plant anatomical entity Plant Compara Genomic alignments Gene tree	Go to Region in Det	ail for more tracks and navigation options (e.g. zooming)		
 Gene gain/loss tree Orthologues 	🌣 🌆 < 🖃 %			
Pan-taxonomic Compara		0.000 M		Forward strand
- Gene Tree	TAIR gene	MD 6.280 MD 6.280 MD	6.200 MD C	6.300 lb 6.3
- Phenotype	Cashing T10520.1 s	110022 1 >	protein	coding protein coding protein coding < AC013354.6
	7 9 11 7 1 7	110022_17		
Genetic Variation	TAIR gene			<atig18210.1< td=""></atig18210.1<>
Genetic Variation Variant table Structural variants	TAIR.gene.	2000 Carlos Carl		<at1g18310.1 protein coding</at1g18310.1

< 🖾 🖸

Agronomic Linked Data (AgroLD)

Search 👻 Documentation About Please send us your feedbackt 🚖

Select a sample query and run it. The sample query could be used to modify the parameters accordingly. Alternatively, enter SPARQL code in the quer

Search > SPARQL Query Editor

1 BASE <http://www.southgreen.fr/agrold/

PREFIX vocab:<vocabulary/>

PREFIX graph:<gramene.cyc>

GRAPH graph: {

PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#</p>

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 pathway:<biocyc.pathway/CALVIN-PW

SELECT DISTINCT ?gene ?name ?taxon_nam

?gene vocab:is_agent_in pathway
?gene rdfs:label ?name.

?gene vocab:taxon ?taxon_name.

The SPARQL Query Editor provides an interactive environment to formulate SPARQL queries.

	Select a type*	•	Type here					Search
PATHWAY :	CALVIN-PWY / Calvir	ı cycle						
URI: <u>http://www.sout</u>	thgreen.fr/agrold/biocyc.pathway/	CALVIN-P	WY					
Participating ge	nes Next page>>							
K3 ±							Search:	Show 30
geneld	\$	gene_n	ame	♦ taxon	:	taxon_name	₽	URI
1 AT1G18270 (display)		fructose	bisphosphate aldolase	http://purl.c (in Sparql)	bolibrary.org/obo/NCBITaxon_370	² Arabidopsis thaliana		http://identifiers.org/ensembl.plant/AT10 (in Sparql)
2 AT1G42970 (display)		glyceraldehyde-3-phosphate dehydrogenase		http://purl.c	bolibrary.org/obo/NCBITaxon_370	² Arabidopsis thaliana		http://identifiers.org/ensembl.plant/AT10 (in Sparql)
3 AT1G43670 (display)		fructose	1,6-bisphosphatase	http://purl.c (in Sparql)	bolibrary.org/obo/NCBITaxon_370	² Arabidopsis thaliana		http://identifiers.org/ensembl.plant/AT10 (in Sparql)
Search > A	dvanced form-ba	sed s	earch					
			Search examples: ontological cor	icepts - 'plant h	eight' or 'regulation of gene expres	sion'; gene names - 'GRP2' or 'TCP2'.		
				QTL ID: '	AQAA003' ; protein name: 'TBP1'			
	Select a type*	-	Type here					Search

The Advanced Search query form is based on the REST API suite, developed under the AgroLD project. The aim of this effort is to provide non-technical users with a tool to query the knowledge base.

We are open to discussions and collaborations. Feel free to get in touch: **Dr. Pierre Larmande:** pierre.larmande@ird.fr **Dr. Manuel Ruiz:** manuel.ruiz@cirad.fr **Mr.Nordine El Hassouni:** nordine.el_hassouni@cirad.fr 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. GreenPhyIDB: 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 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).