AgroLD: A knowledge graph for the plant sciences

Pierre Larmande, Konstantin Todorov

AgroLD: a Knowledge Graph for the Plant Sciences

Pierre Larmande[1,2,3][0000−0002−2923−9790] and Konstantin Todorov[3][0000−0002−9116−6692]

[1] DIADE, IRD, Univ. Montpellier, CIRAD, Montpellier, France
[2] French Institute of Bioinformatics (IFB)—South Green Bioinformatics Platform, Bioversity, CIRAD, INRAE, IRD, Montpellier, France
pierre.larmande@ird.fr
[3] LIRMM, CNRS, Univ. Montpellier, France
todorov@lirmm.fr

Abstract. Recent advances in sequencing technologies and high-throughput phenotyping have revolutionized the analysis in the field of the plant sciences. However, there is an urgent need to effectively integrate and assimilate complementary information to understand the biological system in its entirety. We have developed AgroLD, a knowledge graph that exploits Semantic Web technologies to integrate information on plant species and in this way facilitate the formulation and validation of new scientific hypotheses. AgroLD contains around 900M triples created by annotating and integrating more than 100 datasets coming from 15 data sources. Our objective is to offer a domain specific knowledge platform to answer complex biological and plant sciences questions related to the implication of genes in, for instance, plant disease resistance or adaptative responses to climate change. In this paper, we present results of the project, which focused on genomics, proteomics and phenomics. We present the AgroLD pipeline for lifting the data, the open source tools developed for these purposes, as well as the web application allowing to explore the data.

Keywords: Knowledge Graph · Linked Data · Plant Sciences.

1 Introduction

The understanding of genotype-phenotype interactions, which stand for the regulation of gene expression conferring a phenotype, is one of the most critical research areas in agronomy. However, these interactions are complex to identify because they are expressed at different molecular levels in the plant and are strongly influenced by environmental factors. The new challenges consist in identifying these interactions between various molecular entities involved in the expression of the phenotype, which, we believe, can only be addressed by integrating information from different levels in a global model using a systemic approach in order to understand the real functioning of the biological system.

Recent high-throughput technologies such as Next Generation Sequencing, which allow DNA to be sequenced much more rapidly than previous methods, can only partially capture these dynamics of interactions [1]. Similarly, high-throughput phenotyping, which allows to produce a large amount of experimental data in various environmental conditions, lack in filling the gap with genomics data because of missing links in the data. Even if these new technologies allow to go further and further in obtaining new data, the current limitations and challenges are mainly at the level of data integration and data analysis. Moreover, a methodology to standardize and share data according to the FAIR (Findable, Accessible, Interoperable, Reusable) principles should allow to group data efficiently, and thus contribute to the improvement of the biological knowledge [2]. Indeed, it appears that this knowledge is still fragmented and this fragmentation hinders the elucidation of the molecular mechanisms that govern the expression of complex phenotypes [3].

The question is, therefore, how to structure and manage the complexity of biological data in order to extract knowledge that can be used to identify the molecular mechanisms controlling the expression of plant phenotypes. Our hypothesis is that weaving these data and disparate information into a knowledge graph (KG) would enable the formulation and validation of research hypotheses that would link genotype to phenotype, hence unlocking the potential of the currently available decentralized scientific data.

We have developed AgroLD (for Agronomy Linked Data) [4],[4] a FAIR knowledge graph powered by Semantic Web technologies as a structure to integrate data, to enable knowledge sharing and to allow information retrieval at scale. It is designed to integrate available information on various plant model species in the agronomic domain such as rice, arabidopsis and wheat, to name a few. The online documentation [5] shows the complete list of species with the total number of related protein entities. Among the contributions of the project is the development of the AgroLD schema, which combines newly created concepts/properties with concepts/properties imported from various ontologies from the biology field. Because life sciences and bioinformatics produce a large plethora of specific data formats, specific open source tools for data conversion to the Resource Description Framework (RDF) following the AgroLD schema have been developed and discusses in this paper. These different steps have led to the construction of several graphs on plant molecular interactions, which have been interlinked to form the AgroLD KG. We present these tools, together with a data fusion approach that allows for the construction of the pivotal AgroLD graph. Finally, we introduce an exploratory search engine that allows to browse the knowledge graph.

[4] www.agrold.org
[5] http://www.agrold.org/documentation.jsp

2 Related Work

In the last decade, many initiatives emerged in the biomedical and bioinformatics fields aiming at providing integrated environments to formulate scientific hypotheses about the role of genes in the expression of phenotypes or the emergence of diseases. Among them, we cite Bio2RDF [27], EBI RDF [28], Uniprot RDF [29], WikiPathways [30], OpenPhacts [31] and PubChemRDF [32]. Moreover, we can mention the BioHackathon[6] [33] which gather multidisciplinary scientists to solve biomedical and bioinformatics issues in data integration and knowledge representation. Since 15 years, the BioHackathon produces tools, ontologies [34] and guidelines [35] for RDF modelling and conversion. Recently, the DisGeNET [36] RDF platform and the Monarch Initiative [37] were created for human biology data. OntoForce [7] developed a new tool named DISQOVER for data discovery in life sciences. However, to the best of our knowledge, there was no equivalent in the plant sciences field before the AgroLD platform [4] was launched in 2015. In a related topic, KNETMINER [38] is a graph database for plant molecular network that has been developed with Neo4J and provides also a subset of its datasets through a SPARQL endpoint. Both KNETMINER and AgroLD have the same purpose and target the same scientific community. However, KNETMINER offers limited access to its features in its free version, while AgroLD has the advantage of being open and FAIR.

3 The AgroLD Knowledge Graph

3.1 Overview

AgroLD is built in phases spanning vast aspects of plant molecular interactions. The current phase (second phase) covers information on genes, proteins, predictions of homologous genes, metabolic pathways, plant phenotype and genetic studies. At this stage, we have integrated data from several resources such as Ensembl plants [5], UniProtKB [6], Gene Ontology Annotation [7]. The choice of these sources has been guided by the biological community, as they are widely used and have a strong impact on the user's confidence. We have also integrated resources developed by the local SouthGreen platform [8] such as TropGeneDB [9], a tropical plant genetics database, OryGenesDB [10], a rice genomics database, GreenPhylDB [11], a comparative genomics database for tropical plants, OryzaTagLine [12], a rice phenotype database and SniPlay [23], a rice genomic variation database. These resources bring together experimental data produced by researcher groups in Montpellier and the South of France. The online documentation provides an overview of the integrated data sources [8].
The conceptual framework of AgroLD is based on well-established ontologies in the plant field such as Gene Ontology [14], Plant Ontology [15] or Plant Trait

[6] http://www.biohackathon.org
[7] https://www.ontoforce.com
[8] http://www.agrold.org/documentation.jsp

Ontology [16]. Furthermore, we developed a dedicated schema [9] that creates links between the imported ontologies and introduces new classes and properties. The online documentation [10] shows the complete list of the used ontologies. The majority of these ontologies are hosted by the OBO Foundry project [17].

In the following, we describe the components of the knowledge graph and the process of its construction.

3.2 Statistics

As of today, AgroLD contains more than 900 Millions triples resulting of the integration of roughly 100 datasets gathered in 33 named graphs. Table 1 gives a summary of all number of features. Table 3 gives an overview of available resources and tools. All datasets are available in Zenodo under the Creative Commons Attribution 4.0 International license (CC-BY 4.0). Each resource can contain several datasets, for instances, one dataset per species or per data type. Combining all ontologies and datasets imported, AgroLD graph gather 383 classes and 793 properties. Among the pipelines developed to lift up the datasets, we focused also on connecting our datasets with others. The property *rdfs:seeAlso* reach the total number of almost 80 millions of outbound links making the AgroLD graph correctly linked with other datasets in the LOD. Besides, we paid attention to increasing the number of semantic annotations with imported ontologies, which increased the number of links between datasets making the overall graph denser. We created more than 14 million semantic links linking entities to ontological classes. Finally, our data linking strategy (see next section) allowed us to create around 160,000 *owl:sameAs* links between entities.

Table 1. Features of the AgroLD knowledge graph.

Features	Number of features
datasets	100
graphs	20
triples	933,663,219
classes	383
properties	793
rdfs:seeAlso	79,696,972
owl:sameAs	166,551
semantic annotations	14,652,812

3.3 AgroLD Integration Pipelines

Our contributions focus, among other things, on the development of various RDF conversion workflows for large agronomic datasets. Although several generic tools

[9] https://github.com/SouthGreenPlatform/AgroLD_ETL/tree/master/model
[10] http://www.agrold.org/documentation.jsp

Data sources	Nb of datasets	File format	Ontology used	Nb of triples
Oryzabase	2	TSV	GO,PO,TO	347 K
GO Associations	2	GAF	GO	6,440 K
Genome Hub	7	GFF	GO, SO	12,233 K
Gramene	6	Custom flat file	All	159 K
Ensembl	34	GFF	All	808,874 K
UniprotKB	2	Uniprot	GO, PO	60,034 K
Oryza Tag Line	2	Custom flat file	PO, TO, CO	282 K
TropGeneDB	2	Custom flat file	PO, TO, CO	20 K
GreenPhylDB	2	Custom flat file	GO, PO	3,627 K
SNiPlay	1	HapMap, VCF	GO	16,204 K
Q-TARO	2	TSV	PO, TO	20 K
MSU	2	Custom flat file	PO, TO	2,068 K
RiceNetDB	6	Custom flat file	PO, TO	5,879 K
RapDB	3	GFF	PO, TO	1,026 K
PlantTftDB	12	Custom flat file	PO, TO	86 K
Interpro	1	Custom flat file	PO, TO	196 K
CEGResources	2	GFF	PO, TO	1,031 K
OBO ontologies	12	OWL		15,131 k
TOTAL	100			**934,342 M**

Table 2. Data sources integrated in AgroLD. Ontologies are referenced as GO = gene ontology, PO = plant ontology, TO = plant trait ontology, EO = plant environnment ontology, SO = sequence ontology, CO = crop ontology (plant specific traits)

exist within the Semantic Web community, including Datalift [21], Tarql [22], RML.io [23], none of them were adapted to take into account the complexity of data formats in the biological domain (e.g. VCF format [18]) or even the complexity of the information they could contain. A simple example illustrates this complexity through the GFF (Generic Feature Format) [19], which represents genomic data in a TSV type format (file with tabs as separators). It contains a column with key = value type information, of variable length and having different information depending on the data source. In this case, the transformation needs to be adapted according to the data source. Furthermore, the large volume of data was a limiting factor for the above-mentioned tools. In this context, we developed RDF conversion tools adapted to a large range of genomics data standards such as GFF [19], Gene Ontology Annotation File (GAF) [24], Variant Call Format [18] and we are currently working on packaging these ETL tools in an API [11]. These data standards represent a first step, as they are indeed the most widely used in the community. However, we plan to develop more tools as we will integrate new data standards.

[11] https://github.com/SouthGreenPlatform/AgroLD_ETL

Table 3. Links to AgroLD resource and tools

Name of resource or tool and description, URL
Data
AgroLD datasets, https://doi.org/10.5281/zenodo.4694518
List of graphs, http://www.agrold.org/documentation.jsp
List of ontologies, http://www.agrold.org/documentation.jsp
AgroLD vocabulary, https://github.com/SouthGreenPlatform/AgroLD_ETL/tree/master/model
AgroLD SPARQL Endpoint, http://agrold.southgreen.fr/sparql
Example queries, http://www.agrold.org/sparqleditor.jsp
Use case queries, https://github.com/pierrelarmande/ISWC-use-case
Tools
Web application, https://github.com/SouthGreenPlatform/AgroLD_webapp
RDF conversion pipelines (GFF2RDF, GAF2RDF, VCF2RDF, Datasets), https://github.com/SouthGreenPlatform/AgroLD_ETL

3.4 The AgroLD Schema

In order to match the different data types and properties, we developed a schema that associates the classes and properties identified in AgroLD with corresponding ontologies. Fig. 1 shows an overview of the AgroLD ontology including these mappings. For instance, the *Protein class*[12] is associated with the *SO polypeptide class*[13] with the *owl:equivalentClass* property. Similar mappings have been done for the properties. For example, the *has_function* property is linked with properties from the *RO ontology*,[14] with *owl:equivalentProperty*. When an equivalent property did not exist, we associated it with the higher level property with *rdfs:subPropertyOf*. For example, the property *has_trait*[15], linking entities with TO terms is associated with a more generic property from RO: *causally related to*[16]. So far, 55 mappings have been manually identified.

3.5 URI design

In the transformation pipelines, RDF graphs share a common namespace and are named according to the corresponding data sources. Entities in RDF graphs are linked by the common URI principle. In general, we build URIs by referring to Identifiers.org [19] which provides design patterns for each registered source. For instance, genes integrated from Ensembl Plants are identified by the base URI.[17] When they are not provided by Identifiers.org, new URIs are constructed and in this case URIs take the form.[18] In addition, the properties linking the

[12] http://www.southgreen.fr/agrold/vocabulary/Protein
[13] http://purl.obolibrary.org/obo/SO_000010
[14] http://purl.obolibrary.org/obo/RO_0000085
[15] http://www.southgreen.fr/agrold/vocabulary/has_trait
[16] http://purl.obolibrary.org/obo/RO_0002410
[17] http://identifiers.org/ensembl.plant/{Entity_ID}
[18] http://www.southgreen.fr/agrold/resource/{Entity_ID}

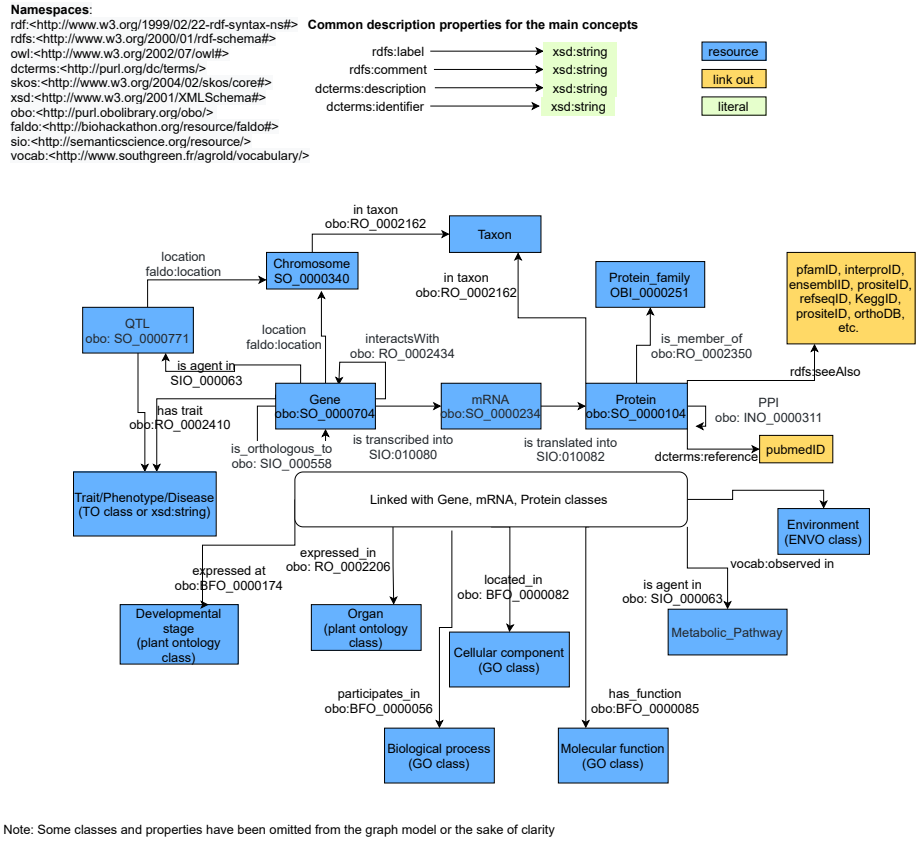

Fig. 1. Overview of the AgroLD schema

entities are constructed as form.[19]

In order to link identical entities from different data sources, we used the approach based on URI pattern matching. Its principle is to scan the URIs in order to look for similar patterns in the terminal part of the URI (i.e. Entity_ID). In addition, we also followed the common URI approach which recommends to use the same URI pattern for two identical entities. Therefore, for the same entity, this allowed us to aggregate information from different RDF graphs. In addition, we used cross-reference links by transforming them to URIs and linking the resource to the *rdfs* predicate `seeAlso`. This significantly increases the number of outbound links by reaching almost 80 million links, making AgroLD better integrated with other data sources. In the future, we plan to implement a similarity entity profile approach to identify matches between entities with different URIs.

[19] http://www.southgreen.fr/agrold/vocabulary/{property}

4 Challenges in Creating the AgroLD Graph

The process of creating a knowledge graph is complex and challenging. In this section, we will present some of the challenges we had to address and in particular those related to managing the heterogeneity of the datasets and their sizes, aligning the entities and assessing the data quality.

Concerning **data heterogeneity**, the main problem was the variety of the data formats which we solved by having RDF as unified format. We proposed several pipelines that were able to handle this variety and manage the size of the datasets. Indeed, as discussed in Section 3.3, in the majority of cases, we preferred developing our own solutions instead of using generic tools to manage better the complexity or the size of the datasets. Another problem was the heterogeneity of the genomic coordinates (i.e. different naming of the chromosome identifier, missing information, etc.). We solved it by choosing a unique representation and transforming all coordinates in URIs patterns following the FALDO ontology representation [34].

Concerning the **entity linking** problem (i.e. same entities having different names or identifiers), we managed to only partially solve this problem, by using pattern matching in URIs, or database cross-links to identify mappings between entities. Indeed, in the case where entities have a different namespace URI (e.g. namespace1:identifier1 and namespace2:identifier1), we search patterns matching in the URIs and create a new URI doing the mapping between them. In the case when entities have different URIs with no matching patterns but having synonym properties (i.e. skos:altLabel, skos:prefLabel, skos:synonym or specific ones), we search matches with these properties and the URIs patterns. For entities that do not contain the above information, we adopt a more global approach based on properties and values analysis. However it is an open challenge that we are currently working on.

Concerning the processes followed for **data quality assessment**, pre-processing quality assessments such as input file format, raw line and missing value check were developed for the resources used by the ETL pipeline. Then, the produced triples were validated for syntax with built-in libraries (e.g. with RDFlib). Further assessments include counting the number of entities (e.g. genes, proteins, chromosomes, etc) and checking the presence/absence of properties with sets of SPARQL queries. More complex quality assessment such as type restriction on properties is planned in the future.

5 Data Access and Applications

The AgroLD KG is available for access via a SPARQL enpoint [20]. However, although the SPARQL language is efficient to build queries, regarding access to

[20] http://agrold.southgreen.fr/sparql

RDF data, it remains difficult to handle for our main users, which are mainly biologists with little or no background in formal query languages. Therefore, we propose a web application implementing various elements of a semantic search systems, such as pattern-based querying, graphical visualization, information retrieval tools.[21]

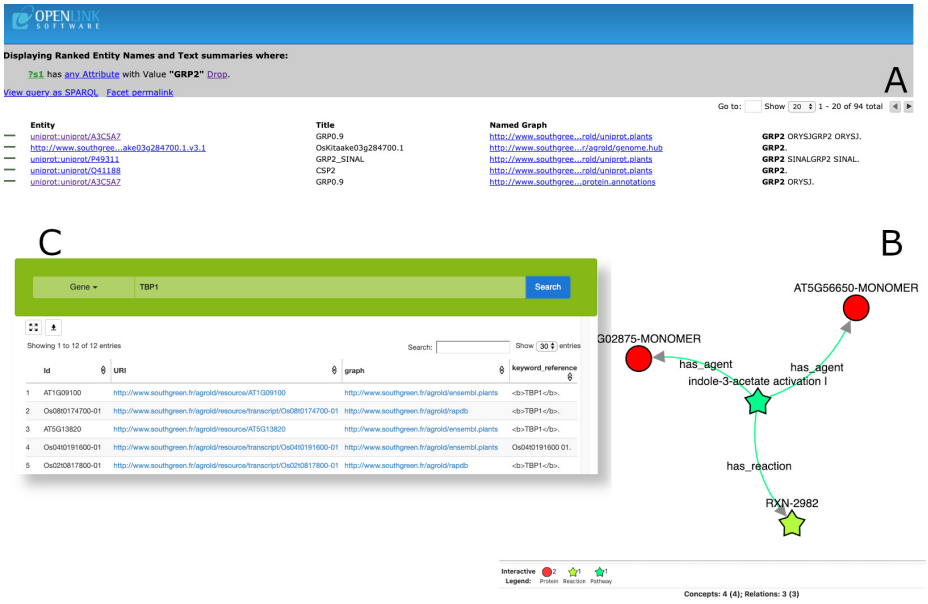

Fig. 2. Overview of AgroLD Web interfaces. (A) displays the Faceted search interface. (B) displays results from the KnetMaps tool. (C) displays results from the advanced search interface.

Hence, the AgroLD platform provides three entry points, as described in [4]:

- *Quick Search* is a faceted search plugin made available by Virtuoso that allows users to search by keywords and browse AgroLD content by navigating through links. Fig. 2A shows the result of a keyword search. In this example a user submitted the *GRP2* keyword which stands for a gene name. Results are ranked according to the number of occurrences found in various fields of the entities.
- *Advanced Search* is an interface allowing specific searches by class of entities such as filtering by Gene, Protein, Pathway and having an aggregation engine for external resources (Fig. 2C). The Advanced Search form is based on a RESTFul API. The purpose of this interface is to provide a tool to query the

[21] http://www.agrold.org

knowledge graph while hiding the technical aspects of SPARQL querying. The interest of coupling the AgroLD RESTFul API and this interface is to be able to interactively combine searches with external services API such as for instance Pubmed or EMBL APIs. Moreover, this is made possible through the user interface and also programmatically. As shown in Fig. 2C, the users begin by entering one or several keywords and select which type of entity they want to retrieve. In this example, we selected the type *Gene* and the keyword *TBP1*. The results are presented in the form of a table which can be sorted and explored. Moreover, results could be visualised as a graph as shown in Fig. 2B. This tool was adapted from KnetMaps [39].

– *The SPARQL Editor* is a query editor that provides an interactive environment for formulating SPARQL queries. We developed the editor based on the YASQE and YASR [26] tools and adapted them for our system. In addition, we proposed a list of modular and customizable query patterns according to the users' needs that can be automatically executed through the editor. Fig. 3 shows the SPARQL Editor. The editor is divided in three areas. The main area on the top left corresponds to the query area. Thanks to the YASQE tool, proposing several code editing functions, the code syntax is highlighted and checked for errors. Furthermore, users can load their own queries stored in a file in order to run batch queries and save their results in various file formats. It is also possible to build up queries by using query patterns. In this case, the users can select one of the dozen query patterns shown on the right. The query appears in the left-hand area within the query text box. Users can read the code of the query and see differently colored pieces of code according to the type of variable, string fields and SPARQL syntax. Hence, users can directly modify the code. There is also a text box above the query box that allows users to modify the value of the string parameter and by clicking on the *apply* button, it modifies the string value of the query. Finally, the results are displayed at the bottom of the editor as a table (by default), but they can also be displayed in JSON or graph-based formats. Each column can be sorted and text filters can be applied to search among the results. Data can be downloaded as a CSV file.

6 Use-case Scenarios

A better understanding of genotype-phenotype relationships requires the integration of biological information of various kinds. However, this information is often dispersed in several databases on the Internet each with different data models, scales or distinct means of access. For biologists, it is difficult to search relevant information in these databases as the mass of information can be incomplete and hard to manage. These problems are particularly relevant in the context of genetic association analyses or GWAS (Genome Wide Association Studies), which allow to associate large regions of the genome (locus) with a phenotypic trait (trait). GWAS loci often include several hundred genes that

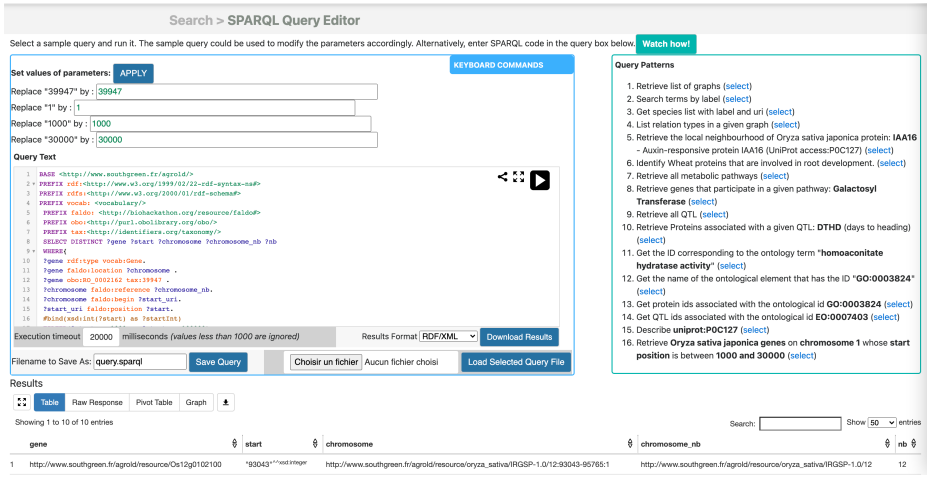

Fig. 3. The SPARQL query editor. The Query patterns frame allows to select a query from a natural language question. The Query text frame allows to visualize and modify the SPARQL query. The results frame displays results returned from the query.

need to be analysed in order to identify only a fraction of the genes associated with the trait under study. At some point, each scientist will have to choose which genes to investigate further in the laboratory.
In order to show how AgroLD can help in this type of analysis, we took the results published in [40] and tried to reproduce the experiments. The paper studies the key genes that are responsible of the panicle architecture in rice. The authors outlined, based on a manual literature review, a list of 319 candidate genes known to regulate the plant architecture. The aim of our use-case study is to reproduce these results automatically.

The authors of [40] identified numerous GWAS loci combining several trait associations all along the chromosomes and studied chromosome 4, which was associated with ten panicle and yield traits. We found less associations with the query Q1 [22] in AgroLD. Indeed, only five phenotypic traits loci associated to "panicle" trait name were retrieved.

Next, the authors identified 20 candidate genes distributed along chromosome 4. By building a second query Q2, which retrieve the genes available for chromosome 4 and using a filter on "panicle", we obtained 15 genes results.

Finally, the authors narrowed down the genomic region of chromosome 4 between 30 Megabases and 32 Megabases. They identified five candidates genes namely OsKS3, OsKS1, OsKS2, OsMADS31 and NAL1. On our side, querying

[22] Example Use case, available from: https://github.com/pierrelarmande/ISWC-use-case

the same genomic region with similar filters (Q3), we obtained only one gene: NAL1. However, using a less restrictive query (Q4), we obtained 81 genes results including the five candidates identified by the authors.

By comparing our results to the ones in the paper, we can first argue that the authors have a larger GWAS/QTL datasets than AgroLD currently has integrated. Thus, they get more genomic regions associated with phenotypic trait than we get in Q1. Second, they have a better selection of candidate genes for a given genomic region. Even if AgroLD contains a large number of genes (81 genes) for the same genomic region, the final result is smaller when genes are filtered by the name of the trait. After checking, we observed that this value is absent in the majority of cases. The authors extracted this information from the manual review of scientific papers.

7 Conclusion and Future Work

Data in the agronomic domain are highly heterogeneous and dispersed. For plant scientists to make informed decisions in their daily work it is critical to integrate information at different scales. Semantic Web technologies play a pivotal role in data integration and knowledge management. The biomedical domain provides a good example to follow by capitalizing on previous experience and addressing lessons learned. To build on this line of research in the agronomy field, we have developed the AgroLD KG. The knowledge base exploits the power of seamless data integration offered by RDF. It contains more than 900 Millions triples resulting of the integration of roughly 100 datasets gathered in 33 named graphs. However, it coverage with respect to the species and the data sources are expected to grow with the subsequent releases. To our knowledge, AgroLD is one of the first initiatives taken to bring Semantic Web practices to the agronomic domain, playing a complimentary role in the integrative approaches adopted by the community.

AgroLD is being actively developed based on feedback from domain experts. It also benefits from the support of the SouthGreen Bioinformatics Platform since its beginning in 2015 by providing IT support and infrastructure to host data and web applications. SouthGreen is one of the core platforms of the French Elixir-EU node, thus will provide a long lasting support for AgroLD. AgroLD is strongly linked to several use-cases of the D2KAB project[23] (National Research Agency funded project) to demonstrate the benefits of linked data to discover gene-phenotype interactions. With the achievement of the second phase, user feedback reveals some limitations and challenges on the current version. Thus, a number of issues are a matter of ongoing or future work.

On the one hand, the KG coverage has to be extended to a larger number of biological entities (e.g. miRNA) and relations (e.g. co-expression, regulation and

[23] https://d2kab.mystrikingly.com/

interaction networks) in order to capture a broader view of the molecular interactions. For instance, we need to integrate information on gene expression and gene regulatory networks. On the other hand, the ETL process for KG creation is mostly based on domain specific approaches thus limiting its re-usability. We will investigate approaches using declarative functions for its creation.

Methods for knowledge augmentation need to be applied and adapted to our data. Indeed, we have observed that certain information remains hidden in the RDF literal contents, such as biological entities or relationships between them, while a wealth of related knowledge is available in external sources. We currently developing methods to extract information embedded in unstructured data such as the text fields from the KG or from external web documents and scientific publications and bring this information under a structured form to the knowledge base. Finally, we are in the process of extending state-of-art data linking techniques by considering the specificities of the biological domain.

References

1. Kemble H, Nghe P and Tenaillon O. Recent insights into the genotype–phenotype relationship from massively parallel genetic assays. Evol Appl. 2019 Oct; 12(9): 1721–1742.
2. Wilkinson MD, Dumontier M, Aalbersberg IjJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3.
3. Weighill D et al. Multi-Phenotype Association Decomposition: Unraveling Complex Gene-Phenotype Relationships. Frontiers in Genetics.2019;10p417. https://doi.org/10.3389/fgene.2019.00417
4. Venkatesan A, Tagny Ngompe G, Hassouni NE, Chentli I, Guignon V, Jonquet C, et al. Agronomic Linked Data (AgroLD): a Knowledge-based System to Enable Integrative Biology in Agronomy. PLoS ONE. 2018;13:17.
5. Bolser D, Staines DM, Pritchard E, Kersey P. Ensembl Plants: Integrating Tools for Visualizing, Mining, and Analyzing Plant Genomics Data. Methods Mol Biol Clifton NJ. 2016;1374:115–40.
6. The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res. 2018;47:D506–15.
7. Huntley RP, Sawford T, Mutowo-Meullenet P, Shypitsyna A, Bonilla C, Martin MJ, et al. The GOA database: gene Ontology annotation updates for 2015. Nucleic Acids Res. 2015;43:D1057-1063.
8. South Green collaborators. The South Green portal: a comprehensive resource for tropical and Mediterranean crop genomics South Green collaborators. Curr Plant Biol. 2016;78:6–9.
9. Hamelin C, Sempere G, Jouffe V, Ruiz M. TropGeneDB, the multi-tropical crop information system updated and extended. Nucleic Acids Res. 2013;41.
10. Droc G, Périn C, Fromentin S, Larmande P. OryGenesDB 2008 update: database interoperability for functional genomics of rice. Nucleic Acids Res. 2009;37:D992-995.
11. Valentin G, Abdel T, Gaëtan D, Jean-François D, Matthieu C, Mathieu R. GreenPhylDB v5: a comparative pangenomic database for plant genomes. Nucleic Acids Res. 2020;

12. Larmande P, Gay C, Lorieux M, Périn C, Bouniol M, Droc G, et al. Oryza Tag Line, a phenotypic mutant database for the Genoplante rice insertion line library. Nucleic Acids Res. 2008;36:D1022-1027.
13. Dereeper A, Homa F, Andres G, Sempere G, Sarah G, Hueber Y, et al. SNiPlay3: a web-based application for exploration and large scale analyses of genomic variations. Nucleic Acids Res. 2015;43:W295-300.
14. Gene Ontology Consortium T. The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Res. Oxford Academic; 2019;47:D330–8.
15. Plant T, Consortium O. The Plant Ontology Consortium and plant ontologies. Comp Funct Genomics. 2002;3:137–42.
16. Cooper L, Meier A, Laporte MA, Elser JL, Mungall C, Sinn BT, et al. The Planteome database: An integrated resource for reference ontologies, plant genomics and phenomics. Nucleic Acids Res. 2018;46.
17. Smith B, Ashburner M, Rosse C, Bard J, Bug W, Ceusters W, et al. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotech. 2007;25:1251–5.
18. 1000 Genome project Consortium. Variant Call Format (VCF). http://samtools.github.io/hts-specs/. Last accessed 4 Apr 2021
19. The formal specification of GFF3. http://www.sequenceontology.org. Last accessed 4 Apr 2021
20. Laibe C, Wimalaratne S, Juty N, Le Novère N, Hermjakob H. Identifiers. org: integration tool for heterogeneous datasets. Dils 2014. 2014;14.
21. Scharffe F, Atemezing G, Troncy R, Gandon F, Villata S, Bucher B, et al. Enabling linked data publication with the Datalift platform. AAAI; 2012
22. Tarql: SPARQL for Tables. https://tarql.github.io. Last accessed 4 Apr 2021
23. Dimou A, Sande MV, Colpaert P, Verborgh R, Mannens E, Van De Walle R. RML: A generic language for integrated RDF mappings of heterogeneous data. CEUR Workshop Proc. 2014.
24. The Gene Ontology Consortium. Gene Annotation File (GAF) specification [Internet]. Available from: http://geneontology.org/page/go-annotation-file-format-20. Last accessed 4 Apr 2021
25. Heim P, Hellmann S, Lehmann J, Lohmann S, Stegemann T. RelFinder: Revealing relationships in RDF knowledge bases. Lect Notes Comput Sci Subser Lect Notes Artif Intell Lect Notes Bioinforma. 2009. p. 182–7.
26. Rietveld L, Hoekstra R. The YASGUI Family of SPARQL Clients. Semantic Web J. 2015;
27. Belleau F, Tourigny N, Good B, Morissette J. Bio2RDF: towards a mashup to build bioinformatics knowledge systems. J Biomed Inform. United States. 41 (5): 706–16.
28. Jupp S, Malone J, Bolleman J, Brandizi M, Davies M, Garcia L, et al. The EBI RDF platform: linked open data for the life sciences. Bioinforma Oxf Engl. 2014;1–2.
29. Redaschi N, Consortium U. UniProt in RDF: Tackling Data Integration and Distributed Annotation with the Semantic Web. Nat Preced [Internet]. 2009; Available from: https://doi.org/10.1038/npre.2009.3193.1
30. Waagmeester, A., Kutmon, M., Riutta, A., Miller, R., Willighagen, E. L., Evelo, C. T., Pico, A. R., Jun. 2016. Using the semantic web for rapid integration of WikiPathways with other biological online data resources. PLoS Comput Biol 12 (6), e1004989+
31. Chichester C, Digles D, Siebes R, Loizou A, Groth P, Harland L. Drug discovery FAQs: workflows for answering multidomain drug discovery questions. Drug Discov Today. 2015 Apr;20(4):399-405.

32. Fu G, Batchelor C, Dumontier M, Hastings J, Willighagen E, Bolton E. PubChemRDF: towards the semantic annotation of PubChem compound and substance databases. J Cheminform. 2015 Jul 14;7:34.
33. Aoki-Kinoshita k et al. Implementation of linked data in the life sciences at BioHackathon 2011. J Biomed Semantics. 2015 Jan 7;6:3. doi: 10.1186/2041-1480-6-3. eCollection 2015.
34. Bolleman JT, Mungall CJ et al. FALDO: a semantic standard for describing the location of nucleotide and protein feature annotation. J Biomed Semantics. 2016 Jun 13;7:39. doi: 10.1186/s13326-016-0067-z.
35. DBCLS guidelines for RDFizing databases. https://github.com/dbcls/rdfizing-db-guidelines. Last accessed 4 Apr 2021.
36. Piñero J, Ramírez-Anguita J, Saüch-Pitarch J, Ronzano F, Centeno E, Sanz F, Furlong L. The DisGeNET knowledge platform for disease genomics: 2019 update. Nucl. Acids Res. (2019)
37. Mungall CJ et al. The Monarch Initiative: An integrative data and analytic platform connecting phenotypes to genotypes across species. Nucl. Acids Res. 2019. (48) D704-D715.
38. Hassani-Pak K, Singh A, Brandizi M, Hearnshaw J, Parsons JD, Amberkar S, Phillips AL, Doonan JH and Rawlings C. (2021). KnetMiner: a comprehensive approach for supporting evidence-based gene discovery and complex trait analysis across species. Plant Biotechnology Journal
39. Singh A, Rawlings CJ, Hassani-Pak K. KnetMaps: a BioJS component to visualize biological knowledge networks. F1000Res. 2018 Oct 17;7:1651.
40. Crowell, S., Korniliev, P., Falcão, A. et al. Genome-wide association and high-resolution phenotyping link Oryza sativa panicle traits to numerous trait-specific QTL clusters. Nat Commun 7, 10527 (2016). https://doi.org/10.1038/ncomms10527