Agronomic Linked Data (AgroLD): a Knowledge-based System to Enable Integrative Biology in Agronomy

Aravind Venkatesan[1,3], Gildas Tagny[1,3], Nordine El Hassouni[1,2,8], Imene Chentli[1,3], Valentin Guignon[4,8], Clement Jonquet[1,3], Manuel Ruiz[1,2,8,9] and Pierre Larmande[*1,3,7,8]

[1] Institut de Biologie Computationnelle (IBC), Univ. Montpellier, Montpellier, France.

[2] CIRAD, UMR AGAP, F-34398 Montpellier, France.

[3] LIRMM, Univ. Montpellier & CNRS, Montpellier, France.

[4] Bioversity International, Montpellier, France.

[7] IRD, Univ. Montpellier, DIADE, Montpellier, France.

[8] South Green Bioinformatics Platform, Montpellier, France.

[9] AGAP, Univ Montpellier, CIRAD, INRA, INRIA, Montpellier SupAgro, Montpellier, France

*** Corresponding author**

Email: pierre.larmande@ird.fr

Abstract

Recent advances in high-throughput technologies have resulted in a tremendous increase in the amount of omics data produced in plant science. This increase, in conjunction with the heterogeneity and variability of the data, presents a major challenge to adopt an integrative research approach. We are facing an urgent need to effectively integrate and assimilate complementary datasets to understand the biological system as a whole. The Semantic Web offers technologies for the integration of heterogeneous data and its transformation into explicit knowledge thanks to ontologies. We have developed the Agronomic Linked Data (AgroLD – www.agrold.org), a knowledge-based system relying on Semantic Web technologies and exploiting standard domain ontologies, to integrate data about plant species of high interest for the plant science community e.g., rice, wheat, arabidopsis. We present some integration results of the project, which initially focused on genomics, proteomics and phenomics. AgroLD is now

an RDF knowledge base of 100M triples created by annotating and integrating more than 50 datasets coming from 10 data sources –such as Gramene.org and TropGeneDB– with 10 ontologies –such as the Gene Ontology and Plant Trait Ontology. Our objective is to offer a domain specific knowledge platform to solve complex biological and agronomical questions related to the implication of genes/proteins in, for instances, plant disease resistance or high yield traits. We expect the resolution of these questions to facilitate the formulation of new scientific hypotheses to be validated with a knowledge-oriented approach.

1 Introduction and Background

Agronomy is a multi-disciplinary scientific discipline that includes research areas such as plant molecular biology, physiology and agro-ecology. Agronomic research aims to improve crop production and study the environmental impact on crops. Accordingly, researchers need to understand the implications and interactions of the various biological processes, by linking data at different scales (e.g., genomics, proteomics and phenomics). We are currently witnessing rapid advances in high throughput and information technologies that continue to drive a flood of data and analysis techniques within the domains mentioned above. However, much of this data or information are dispersed across different domain or model specific databases, varied formats and representations e.g., TAIR, GrainGenes and Gramene. Therefore, using these databases more effectively and adopting an integrative approach remains a major challenge.

Among the numerous research directions that the field of bioinformatics has taken, knowledge management has become a major area of research, focused on logically interlinking information and the representation of domain knowledge [1]. To this end, ontologies have become a cornerstone in the representation of biological and more recently agronomical knowledge [2]. Ontologies provide the necessary scaffold to represent and formalize biological concepts and their relationships. Currently, numerous applications exploit the advantages offered by biological ontologies such as: the Gene

Ontology [3] –widely used to annotate genes and their products– Plant Ontology [4], Crop Ontology [5], Environment Ontology [6], to name a few. Ontologies have opened the space to various types of semantic applications [7,8] to data integration [9], and to decision support [10]. Semantic interoperability has been identified as a key issue for agronomy, and the use of ontologies declared a way to address it [11]. Furthermore, efficient knowledge management requires the adoption of effective data integration methodologies. This involves efficient semantic integration of the disparate data sources, making information machine-readable and interoperable. Accordingly, Semantic Web standards and technologies enforced by the W3C, and embracing Tim Berners-Lee's vision [12], offers a solution to facilitate integration and interoperability of highly diverse and distributed data resources. The Semantic Web technologies stack includes among others the following W3C Recommendations: the Resource Description Framework (RDF) [13] as a backbone language to describe resources with triples, RDF Schema (RDFS) [14] to build lightweight data schemas, Web Ontology Language (OWL) [15] to build semantically rich ontologies and the SPARQL Query Language (SPARQL) [16] to query RDF data . All of the previous languages rely on Unique Resource Identifiers (URIs) to define a resource and its components, enabling data interoperability across the Web. RDF describes a resource and its relationships/properties in the form of simple triples, i.e., *Subject-Predicate-Object* offering a very convenient framework for integrating data across multiple platforms assuming the platforms share some common vocabularies to describe their objects. These triples can be combined to construct large networks of information (also known as RDF graphs). A successfully implemented Semantic Web application allows scientists to pose very complex questions through a query or a set of queries that would return highly relevant answers to those questions, facilitating the formulation of research hypotheses [17,18].

There are other approaches to meet the current data integration challenges, e.g., data warehouses. For instance, Intermine [19] has developed a sophisticated application to accommodate the dynamic nature

of biological data and simplify data integration. However, with integrative biology gaining popularity, it is necessary to preserve and share the semantics between the various datasets and make information machine interoperable, enabling large scale analyses of information available over the Web. The Semantic Web approach provides an added value, playing a complementary role to the traditional methods of data integration.

In the recent years, the biomedical community has strongly embraced the Semantic Web vision as demonstrated by a number of initiatives to provide ontologies [20,21] and use them for producing semantically rich data such as in Bio2RDF [22] , OpenPHACTS [23], Linked Life Data [24], KUPKB [25] , and the EBI RDF Platform [26]. In particular, OpenPHACTS serves as a good example of what can be achieved by using Semantic Web knowledge bases. The OpenPHACTS Explorer (http://www.openphacts.org/open-phacts-discovery-platform/explorer) provides use case driven tools that aid in browsing and visualizing the underlying knowledge represented in RDF which is very convenient for biologists.

Currently, there is a growing awareness within the agronomic domain towards efficient data interoperability and integration [2,27,28]. The need for an umbrella approach for providing uniform data is a widely-discussed topic. For instance, the Agriculture Data Interoperability Interest Group (https://rd-alliance.org/groups/agriculture-data-interest-group-igad.html) instituted by the Research Data Alliance (RDA) and agINFRA EU project (www.aginfra.eu) are initiatives that work on improving data standards and promoting data interoperability in agriculture. Moreover, the community has recently also started to adopt AgroPortal [11] as an vocabulary and ontology repository for agronomy –and related domains such as nutrition, plant sciences and biodiversity– that support browsing, searching and visualizing domain relevant ontologies, ontology alignments and creation of semantic annotations. While plant-centric ontologies are now being used to annotate data by various databases [2, 5,28], unlike in the biomedical domain, the adoption of Semantic Web in agronomy is

yet to be completely exploited. Given that agronomic studies involve multiple domains, publicly available knowledge bases such as EBI RDF, Linked Life Data and Bio2RDF serves only limited agronomical information. Hence, it is necessary to build on previous efforts and complete them to provide information compliant with Semantic Web principles within agronomic sciences. This adoption would certainly allow the homogenization of multi-scale information, thereby aiding in the discovery of new knowledge. Therefore, we have developed an RDF knowledge-based system, fully compliant with the Semantic Web vision, called Agronomic Linked Data (AgroLD – www.agrold.org) presented hereafter. The aim of our effort is to provide a portal (to discover) and an endpoint (to query) for integrated agronomic information and to aid domain experts in answering relevant biological questions.

The rest of the paper is organized as follows: in the next section, we describe the data sources integrated or used for the integration, the content and architecture of the knowledge-based system. In the following sections, we present the user interface with some examples queries, then we discuss about the contributions and the future directions.

111 ## 2 Materials and Methods

112 ### 2.1 Information sources

113 AgroLD was conceived to accommodate molecular and phenotypic information available on various
114 plant species (see Fig. 1). The conceptual framework for the knowledge in AgroLD is based on well-
115 established ontologies: GO, SO, PO, Plant Trait Ontology (TO) and Plant Environment Ontology (EO).
116 Among these PO, TO and EO are currently developed by the Planteome project (http://planteome.org).
117 Furthermore, considering the scope of the effort, we decided to build AgroLD in phases. The current
118 phase (phase I) covers information on genes, proteins, ontology associations, homology predictions,
119 metabolic pathways, plant traits, and germplasm, relevant to the selected species. At this stage, we
120 have incorporated the corresponding information from various databases, such as Gramene [29],
121 UniprotKB [30], Gene Ontology Annotation [31], TropGeneDB [32], OryGenesDB [33], Oryza Tag
122 Line [34], GreenPhylDB [35] and SNiPlay [36]. The selection of these data sources was considered
123 based on popularity among domain experts such as GOA, Gramene, and complementary information
124 hosted by the local research community, for instance, Oryza Tag Line and GreenPhylDB. Information
125 on the integrated databases can be found in the documentation page
126 (http://www.agrold.org/documentation.jsp). Table 1 provides a break-down of the data sources and
127 the species covered.

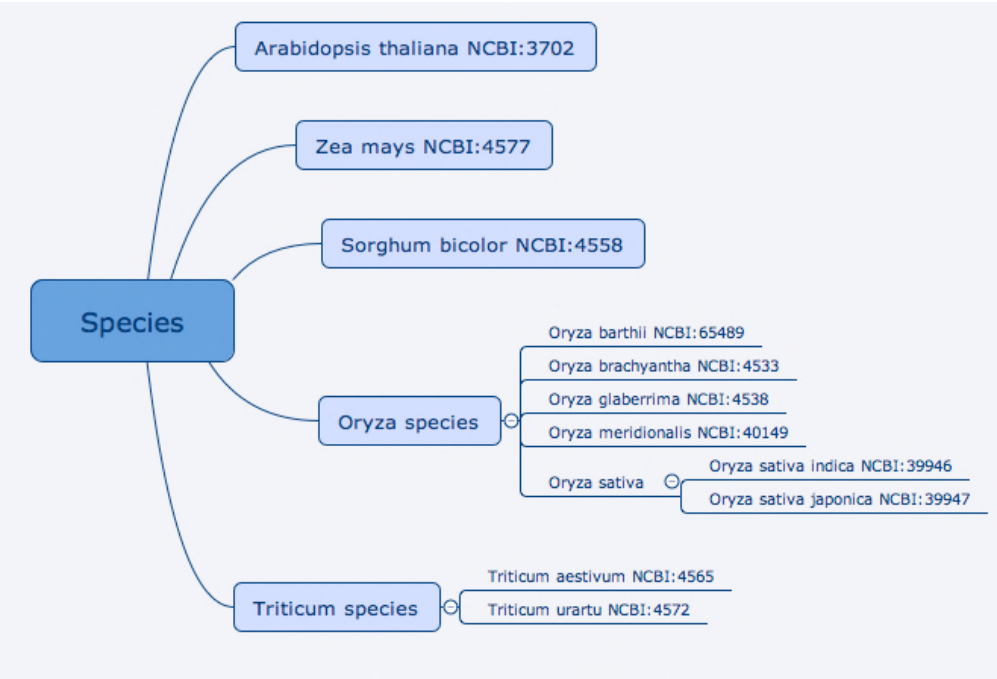

Fig 1. Current plant species included in AgroLD.

Data sources	URL s	File format	# tuples	Crops	Ontologies used	# triples produced
GO associations	geneontology.org	GAF	1, 160K	R, W, A, M, S	GO, PO, TO, EO	2, 700K
Gramene	gramene.org	Custom flat file	1, 718K	R, W, M, A, S	GO, PO, TO, EO	5, 172K
UniprotKB	uniprot.org	Custom flat file	1, 400K	R, W, A, M, S	GO, PO	50, 000 K
OryGenesDB	orygenesdb.cirad.fr	GFF	1, 100K	R, S, A,	GO, SO	2, 300K
Oryza Tag Line	oryzatagline.cirad.fr	Custom flat file	22K	R	PO, TO, CO	300K
TropGeneDB	tropgenedb.cirad.fr	Custom flat file	2k	R	PO, TO, CO	20K
GreenPhylDB	greenphyl.org	Custom flat file	100K	R, A	GO, PO	700K
SNiPlay	sniplay.southgreen.fr	HapMap, VCF	16K	R	GO	16, 000K
Q-TARO	Qtaro.abr.affrc.go.jp	Custom flat file	2K	R	PO,TO	20K
Oryzabase	shigen.nig.ac.jp/rice/oryzabase	Custom flat file	17K	R	GO,PO,TO	153K
TOTAL						87, 400K

Table 1. Plant species and data sources in AgroLD. The number of tuples gives an idea of the number of elements we have annotated from the data sources (e.g., 1160K Gene Ontology annotations). The crops & ontologies are referred as follows: R=rice, W=wheat, A=Arabidopsis, S= sorghum, M= maize, GO = Gene Ontology, PO = Plant Ontology, TO = Plant Trait Ontology, EO = Plant Environment Ontology, SO = Sequence Ontology, CO = Crop Ontology (specific trait ontologies).

2.2 Architecture

AgroLD relies on the RDF and SPARQL technologies for information modelling and retrieval. We use OpenLink Virtuoso (version 7.2) to store and access the RDF graphs. The data from the selected databases were parsed and converted into RDF using a semi-automated pipeline. The pipeline consists of several parsers to handle data in a variety of formats, such as the Gene Ontology Annotation File (GAF) [37], GFF3 [38], HapMap [39] and VCF [40]. Fig. 2 shows the Extraction-Transform-Load (ETL) processes developed to transform in RDF various source data formats. The source code of the ETL workflow is available on GitHub (https://github.com/SouthGreenPlatform/AgroLD).

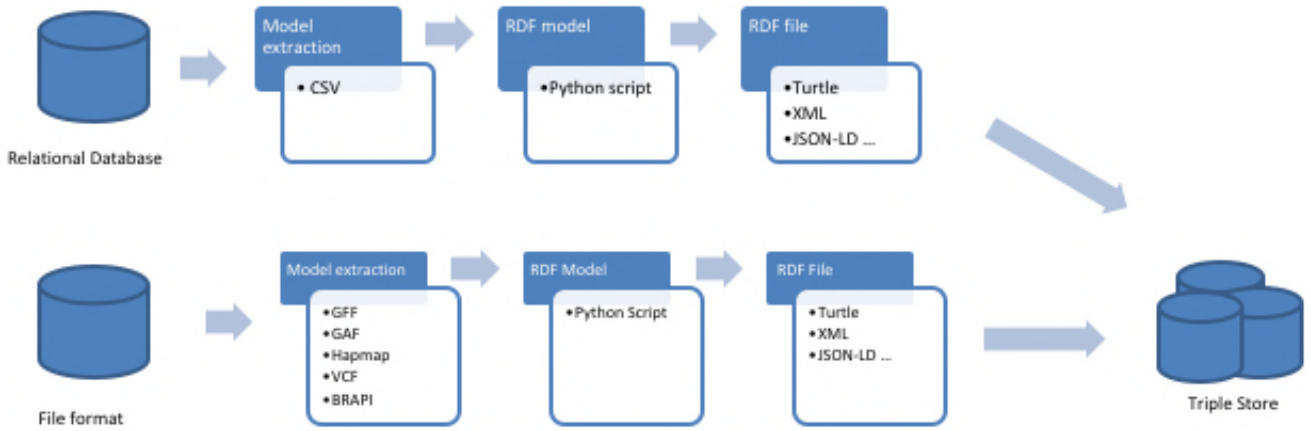

Fig 2. ETL workflow for the various datasets and data formats. The workflow shows two types of process: 1) from relational databases through a CVS file export: in that case, the transformation is tailored for the database model with some Python scripts converters. 2) from standards file formats: in that case, the transformation is generic with some Python packages used as converter tools. The workflow outputs can be produce in various type of RDF format such as turtle, JSON-LD, XML.

For this phase, each dataset was downloaded from curated sources and was annotated with ontology terms URIs by reusing the ontology fields when provided by the original source. Additionally, we use the AgroPortal web service API to retrieve the URI corresponding to the taxon available for some data standards such as GFF. At the end of phase 1, early 2018, the AgroLD knowledge base contains around 100 million RDF triples created by converting more than 50 datasets from 10 data sources. Additionally, when available, we used some semantic annotation already present in the datasets such as, for instances, genes or traits annotated respectively with GO or TO identifiers. In that case, we produce additional properties with the corresponding ontologies thus adding 22% additional triples validated manually (see details in Table 1). The OWL version of the candidate ontologies were directly loaded into the knowledge base but their triples are not count in the total.

160 The RDF graphs are named after the corresponding data sources (protein/qtl ontology annotations
161 being the exception), sharing a common namespace: "http://www.southgreen.fr/agrold/". The entities
162 in the RDF graphs are linked by shared common URIs. As a design principle, we have used URI
163 schemes made available by the sources (e.g., UniprotKB) or by Identifiers.org registry
164 (*http://identifiers.org* - [41]). For instances, proteins from UnitProtKB are identified by the base URI:
165 *http://purl.uniprot.org/uniprot/*; genes incorporated from Gramene/Ensembl plants are identified by
166 the base URI: *http://identifiers.org/ensembl.plant/*. New URIs were minted when not provided by the
167 sources or the by Identifiers.org such as TropGene and OryGenesDB; in such cases the URIs take the
168 form *http://www.southgreen.fr/agrold/**[resource_namespace]/[identifier]***. Furthermore, properties
169 linking the entities took the form: *http://www.southgreen.fr/agrold/vocabulary/**[property]***. An outline
170 of how the RDF graphs are linked is shown in Fig. 3.

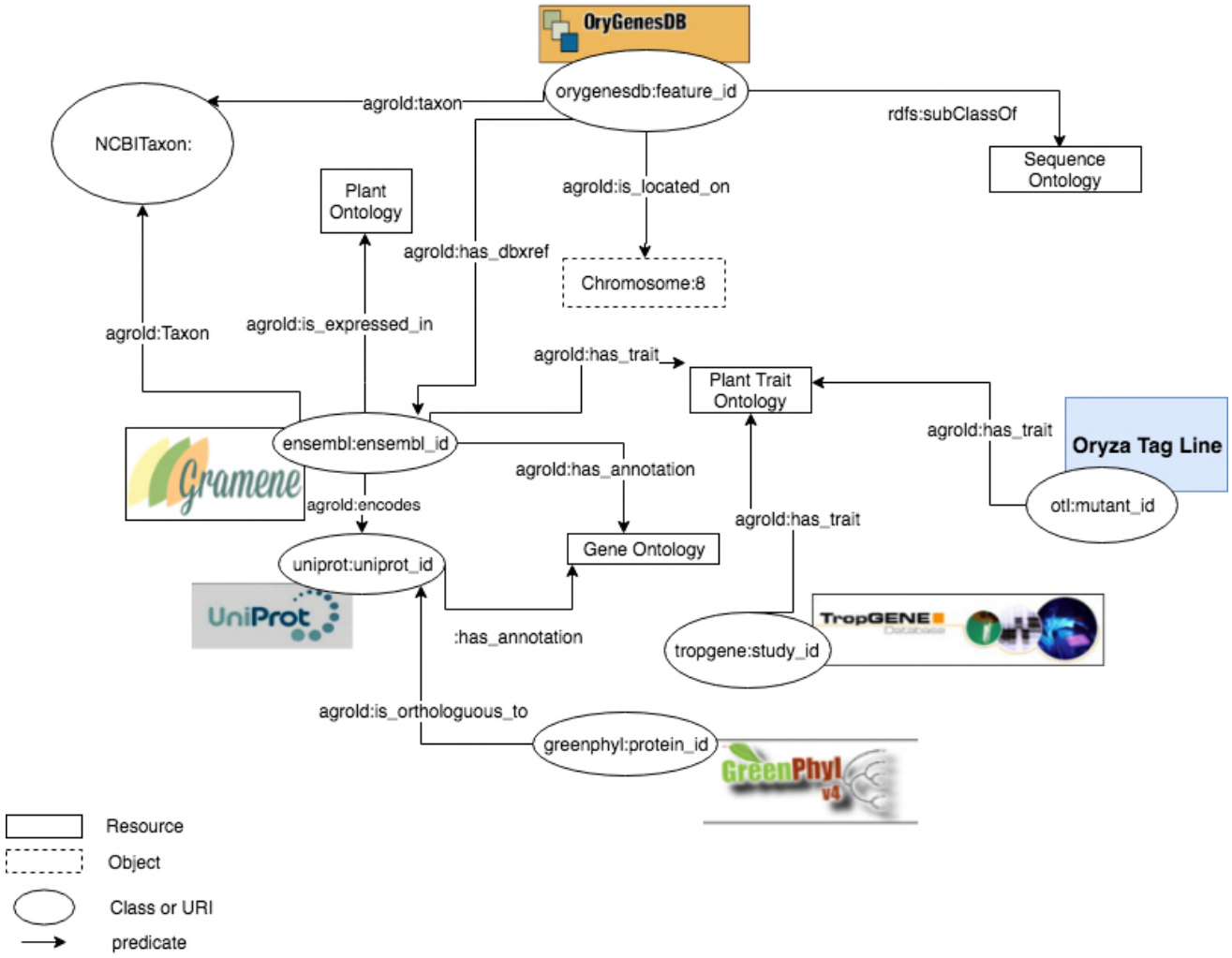

Fig 3. Linking information in AgroLD. The figure illustrates the linking of varies information in AgroLD.

173 To map the various data types and properties, we developed a lightweight schema (cf.
174 https://github.com/SouthGreenPlatform/AgroLD) that glues classes and properties identified in
175 AgroLD and the corresponding external ontologies. For instance, the class Protein
176 (*http://www.southgreen.fr/agrold/resource/Protein*) is mapped as *owl:equivalentClass* to class
177 polypeptide (*http://purl.obolibrary.org/obo/SO_0000104*) from SO. Similar mappings have been made
178 for properties, e.g., proteins/genes are linked to GO molecular function by the property
179 *http://www.southgreen.fr/agrold/vocabulary/has_function*, which is mapped as
180 *owl:equivalentProperty* to the corresponding Basic Formal Ontology (BFO) term
181 (*http://purl.obolibrary.org/obo/BFO_0000085*). When an equivalent property did not exist, we
182 mapped then to the closest upper level property using *rdfs:subPropertyOf* e.g., the property *has_trait*
183 (*http://www.southgreen.fr/agrold/vocabulary/has_trait*), links proteins to TO terms. It is mapped to a
184 more generic property, *causally related to* in the Relations Ontology [42]. For now, 55 mappings were
185 identified. Furthermore, mappings are both stored side by side with ontologies in AgroPortal, which
186 allows direct links between classes and instances of these classes in AgroLD. For example, the
187 following link will show the external mappings for SO:0000104 (polypeptide) stored in AgroPortal:
188 http://agroportal.lirmm.fr/ontologies/SO/?p=classes&conceptid=http%3A%2F%2Fpurl.obolibrary.or
189 g%2Fobo%2FSO_0000104&jump_to_nav=true#mappings. Additionally, classes, properties and
190 resources (e.g., http://www.southgreen.fr/agrold/page/biocyc.pathway/CALVIN-PWY) are
191 dereferenced on a dedicated Pubby server [43]. For details on the graphs, URIs and properties, the
192 reader may refer to AgroLD's documentation (http://www.agrold.org/documentation.jsp).

2.3 User Interface

The AgroLD platform provides four entry points to access the knowledge base:

- *Quick Search (*http://www.agrold.org/quicksearch.jsp), a faceted search plugin made available by Virtuoso, that allows users to search by keywords and browse the AgroLD's content;
- *SPARQL Query Editor (*http://www.agrold.org/sparqleditor.jsp), that provides an interactive environment to formulate SPARQL queries to non-technical users;
- *Explore Relationships* visualizer (http://www.agrold.org/relfinder.jsp), which is an implementation of RelFinder [44] that allows users to explore and visualize existing relationships between entities;
- *Advanced Search* (http://www.agrold.org/advancedSearch.jsp), a query form providing entity (e.g., gene) specific information retrieval.

Alternatively, some user management features have been implemented on the platform. Users have the opportunity to save their search and results on a persistent history session attached to their own account. Furthermore, they can manage search history by editing, deleting or re-running previous searches and exporting results according several formats. In the future, we plan to develop some recommendation features and sharing results between users. More detailed descriptions and figures of the different user interfaces will be provided in the following section.

3 Results and Discussion

RDF knowledge bases are accessed via SPARQL endpoints and in certain cases equipped with faceted browser interfaces. Using SPARQL endpoints require a minimal knowledge of SPARQL, this may result in the resources not being exploited completely. Alternatively, faceted browser interfaces help the user in getting acquainted with information in the resource (e.g., retrieving a local neighborhood for a particular term), the presence non-textual details (e.g., URIs) in the results could be confusing. To this end, we attempted to lower the usability barrier by providing tools to explore the knowledge base. In this section, we demonstrate the complementary role of the *Advanced Search* and *Explore Relationships* query tools with that of the *SPARQL Query Editor*.

We developed the SPARQL Query Editor based on the YASQE and YASR tools [45] and customized for our system. The query page includes a list of modularized example queries, customizable according to the users' needs.

For the comparison, we consider a sample question: '*Retrieving genes that participate in Calvin cycle*'; (Q6 in the online list of modularized queries). As illustrated in Fig 4, the user can run the query to retrieve the list of genes participating in the given pathway (Fig. 4a). Additional information on a gene of interest can be retrieved by clicking on the URI. For example, clicking on AT1GI870 (http://identifiers.org/ensembl.plant/AT1G18270) redirects the users to the gene information provided by Gramene/Ensembl Plants resource (Fig. 4b). The query can be saved and the results can be downloaded in a variety of formats such as JSON, TSV, and RDF/XML. Additionally, user defined queries could also be uploaded.

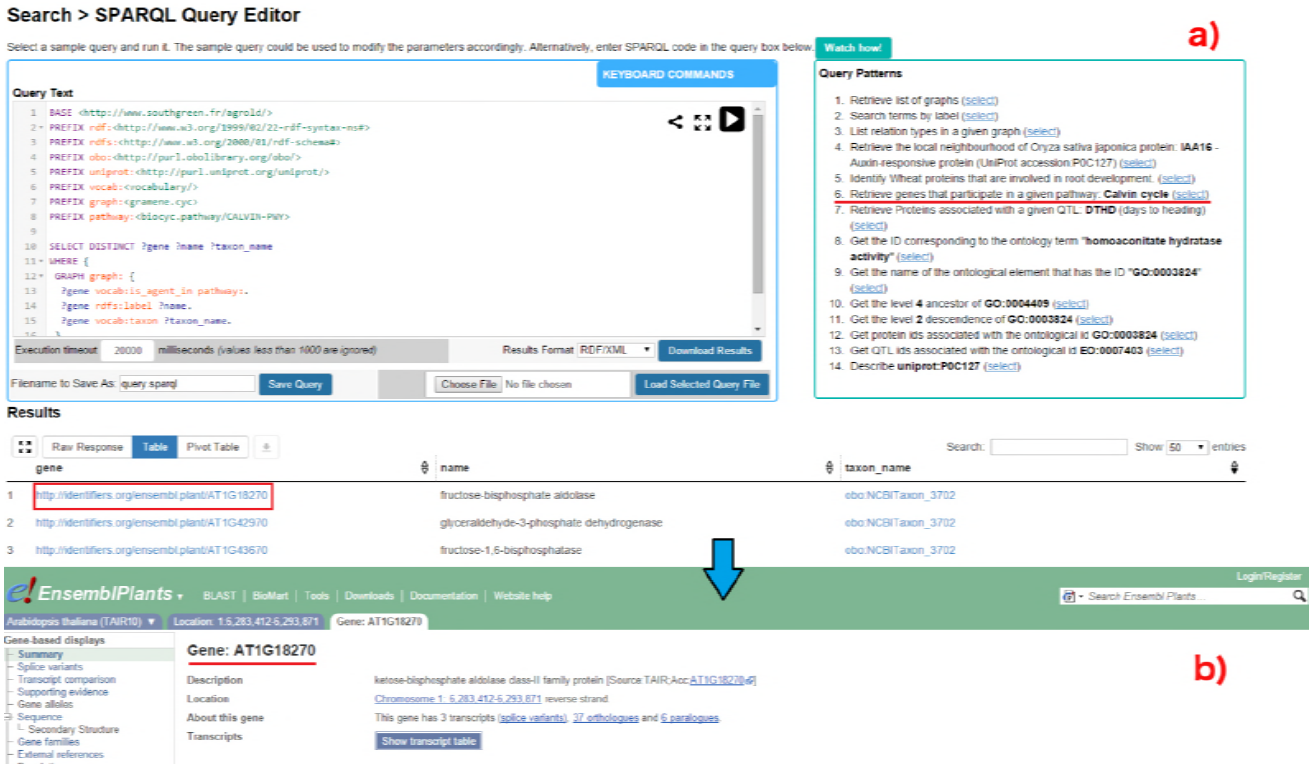

Fig 4. SPARQL Query Editor. Figure illustrates the execution of query Q6: (a) Q6 is one the examples queries on the top-right corner (highlighted in red). On executing the query, the results are rendered below the editor; (b) the user can look up specific genes of interest by clicking on the corresponding URI, which points to the original information source (in this case EsemblPlants).

The *Explore Relationships* tool is based on RelFinder visualization module. This tool aids in visualizing relationships between entities and searching entities by keyword when their URIs are ignored. However, the original version of RelFinder was developed (in ActionScript) and configured for DBpedia. We proposed a configuration and modification of the system suitable for AgroLD. The configuration mainly concerns the SPARQL access point, the properties to be considered for the search of entities and for the description of the resources. Furthermore, we have added some biological examples to guide users. In Fig. 5, the tool is used to search for genes involved in Calvin cycle by entering the name of the entities.

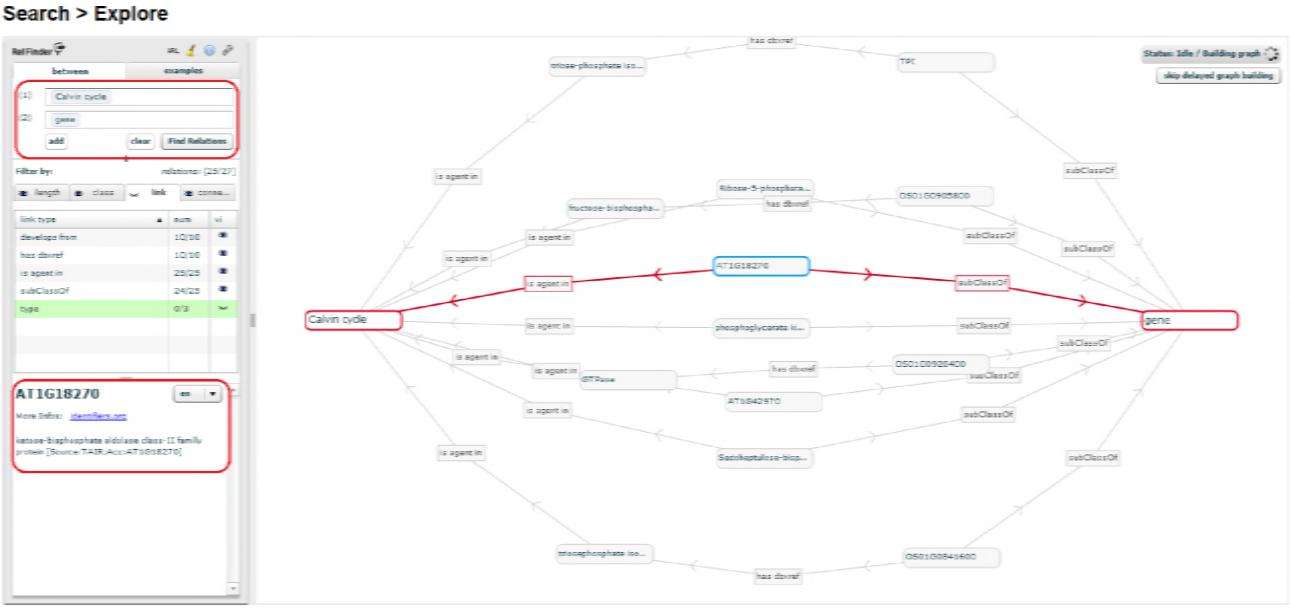

Fig 5. Exploring entity relationships in AgroLD. Figure illustrates differently the results obtained for Q6 using Explore Relationships tool. The results of Q6 can be visualized by entering the concepts (Calvin cycle and gene) in the left panel. On executing the query, all the genes involved in the chosen pathway are revealed. The visualized graph can be altered based on the user interest. Additionally, a gene could be selected (circled on the left) and further explored by clicking on the *More Info* link which directs the user to the information source

The *Advanced Search* query form is based on the REST API suite (http://www.agrold.org/api-doc.jsp), developed completely within the AgroLD project. The aim of this feature is to provide non-technical users with a tool to query the knowledge base. Fig. 6 illustrates steps involved in retrieving information for Q6, using the query form:

a) The user selects *Pathways* from the list of entities and enters the pathway of interest, in this case, Calvin cycle (Fig. 6a);

b) The list of genes involved in the pathway can be retrieved by selecting the pathway.

Furthermore, information on a gene of interest can be retrieved by selecting the specific gene (Fig. 6b). For instance, clicking on AT1GI870 (Fig. 6c) displays all the proteins the gene encodes and the pathways the gene participates in (apart from Calvin cycle). The RESTful API supports the query form and was developed for programmatic retrieval of entity specific knowledge represented in AgroLD. The current version of the API suite (ver. 1) can be used to retrieve gene and protein information, metabolic pathways, and proteins associated with ontological terms. This is achieved by querying entity by name or identifier. A more detailed comparison of the various query form is provided in Supplementary material.

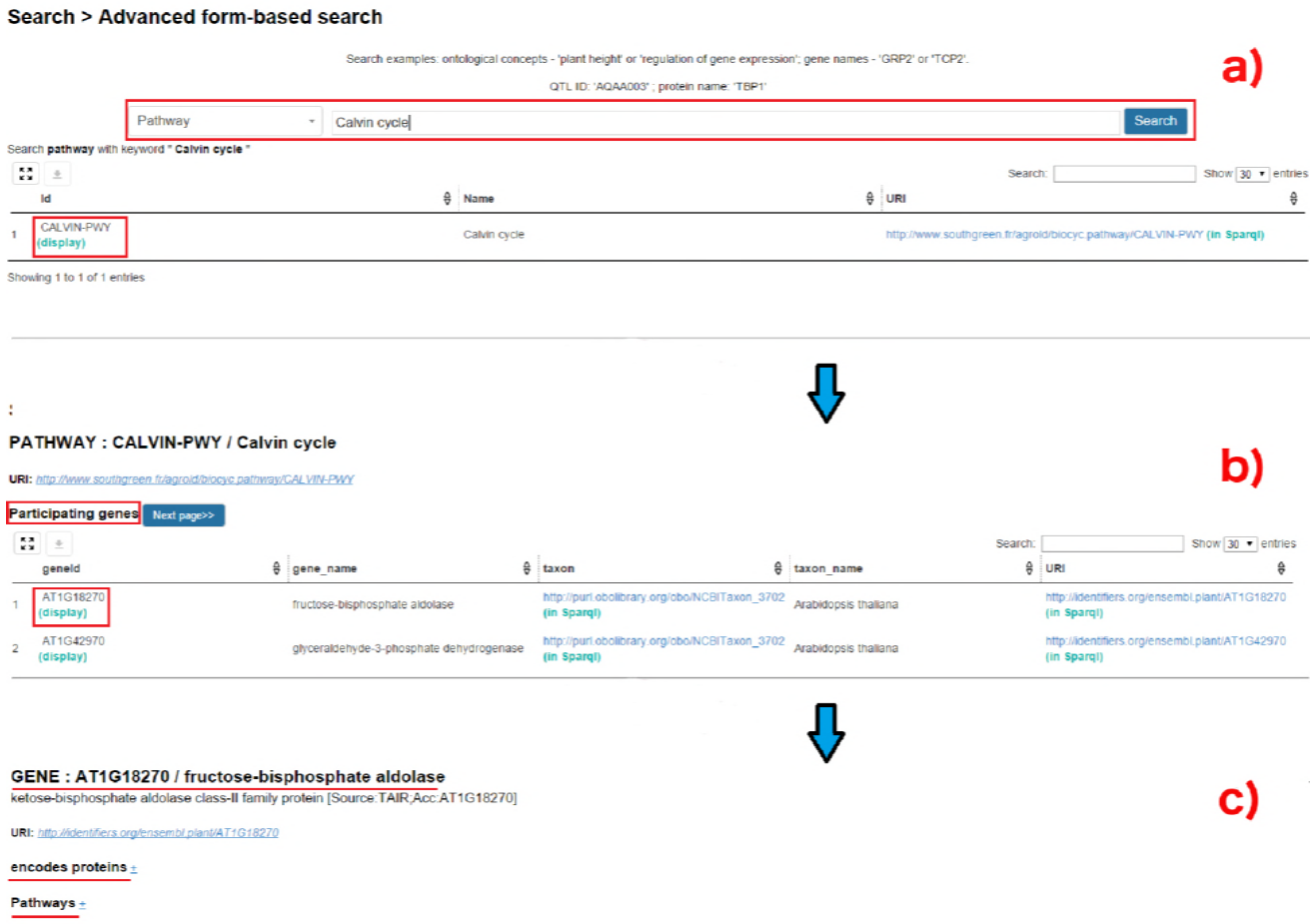

Fig 6. Advanced Search query form: Figure demonstrates the steps involved in retrieving the results for Q6 using the Advanced Search query form: (a) query Q6 can be executed by selecting the type of entity (Pathways – highlighted in red) to search and entering the name of the entity (Calvin cycle). The API then displays the matched results; (b) Clicking on the

result displays the genes participating in Calvin cycle; (c) selecting a gene of interest displays more information pertaining to that gene, for instance, encoding proteins and pathways this selected gene participates in.

4 User Evaluation

AgroLD is being actively developed based on usability testing sessions conducted with domain experts, including doctoral students in biology, curators and senior researchers. Test sessions were designed to measure if:

- Resources integrated in AgroLD are useful;
- AgroLD is easy to use.

For the evaluation of semantic search systems, Elbedweihy et al. [46] recommend a survey of users based on their experience with a few queries submitted to the system. We have used this approach to collect user opinions, comments and suggestions via a feedback form directly within the AgroLD web application. The form includes some questions from the "System Usability Scale" questionnaires [47] and other questions that we considered important. The three main criteria evaluated are:

1. Usability –ease to submit a query (number of attempts, time required) and presentation of the results;
2. Expressiveness – type of queries a user is able to formulate (e.g., keywords or more complex expressions);
3. Performance –speed, correctness and completeness of the results.

In 2017, 8 participants were invited during 2 testing sessions, to search for concepts, genes, or pathways of their interests; and the online form was active (http://agrold.org/survey.jsp) to allow new feedbacks during the exploitation phase. Each question had 5 possible answers ranked from the highest to the lowest note (5 to 1). Some search examples are given in the Supplementary material.

Overall, the idea of data navigation and traversal through knowledge graphs was well received. *Quick Search* won votes for usability and performance (7 participants noted 5), despite several comments to improve the ranking and presentation of results (4 user's comments). *SPARQL Query Editor* won the expressiveness evaluation (5 participants noted 5). *Advanced and Explore search* got average scores but good comments on the capability of discovering unexpected results (e.g., nearest neighbour entities in the graph for the Explore Search and additional results from external Web services for Advanced Search). These user feedbacks reinforced the need for knowledge bases such as AgroLD, wherein users could retrieve information across various data types and sources. This knowledge discovery is supported by the use of shared URI schemes and domain ontologies. The testing sessions also helped us to identify areas for further improvement. Plus, we received suggestions on improving the AgroLD's coverage with more data types such as gene expression data, and protein-protein interactions. Considering, linked data and Semantic Web are still not widely adopted in agronomy, increasing AgroLD's coverage will be an incremental process engaging our user community. This situation is expected to improve with new community efforts such as the Agrisemantics RDA Working Group (https://rd-alliance.org/groups/agrisemantics-wg.html), which role is to reinforce the adoption of semantic technologies in the agri-food domain. We may also mention the AgBioData consortium (https://www.agbiodata.org, [2]) which promotes the FAIR (Findable, Accessible, Interoperable and Reusable) data principles [48] within agricultural research.

Furthermore, we observed that although the information integrated in AgroLD came from curated sources, scientists often prefer to validate these knowledge statements against assertions made in scientific articles. Currently, we have implemented an external Web Services as part of the *Advanced Search Form* to automatically search for publications related to a protein or gene of interest in PubMed Central (cf. Supplementary material and aggregates them within the result of the AgroLD query). However, this feature does not provide detailed (sentence level) assertions described in those

315 publications. This is an area that requires further work. With the recent developments towards making
316 text mined (sentence level) annotations available as RDF [49], query federation can be explored to
317 retrieve entity specific assertions. This would serve as an additional provenance layer.

318 **5 Limits and Perspectives**

319 With the achievement of the first phase of AgroLD, many plant scientists can benefit from the
320 interoperability of the data, but user feedback reveals some limitations and challenges on the current
321 version of AgroLD. In order to achieve the expectations of the scientists for the use of Semantic Web
322 technologies in agronomy, a number of issues need to be addressed:

323 • The coverage content has to be extended to a larger number of biological entities (e.g., miRNA,
324 mRNA) or interaction between them (e.g., co-expression, regulation and interaction networks)
325 in order to capture a broad view of the molecular interactions.
326 • We have observed many information remains hidden in RDF literal contents such as biological
327 entities or relationship between them. This information is poorly annotated (i.e., plain text not
328 formally expressed) and new research methods to identify biological entities and reconstruct
329 their relations further allowing the discovery of relevant links between related resources are
330 required.
331 • The explosion of data in agronomy forces database providers to augment the frequency of their
332 releases. The survey shows a growing interest of using up to date information from the original
333 sources. This have to be taken into account for the updating process in AgroLD.
334 • The user interfaces show some limitations to manage responses with large number of results,
335 e.g., to filter and rank them with precision score.

336 These limitations identified in the current version of AgroLD will be improved in the following
337 versions. We will focus on the following areas:

- User Interface: we plan to explore features offered by Elastic search tool (https://www.elastic.co), to enabling *Quick Search* retrieving retrieve more textual information and hiding the technical details. Further, we will improve the performance and expand the API suite to cover other entities represented in AgroLD (e.g., genomic annotation and homology information).
- Content: integrate information on gene expression such as IC4R [50], Gene Expression Atlas [51], on gene regulatory networks such as RiceNetDB [52] and explore linking text-mined annotations from publications. Support molecular interaction networks per species and also allow knowledge transfer between species.
- Knowledge discovery: explore methods to aid generating hypotheses by retrieving implicit knowledge, e.g., inference rules, automatic data linking, entity recognition, text mining, automatic semantic annotations.
- Data provenance: develop a provenance and annotation model. Set up a validation process to allow users validating computed facts such as semantic annotations automatically produced and attached to a biological entity.
- Updates: To keep AgroLD updated with the latest available data, by processing regular data updates and potentially re-building the entire repository from scratch every 12 months.[1] Additionally, we plan to fully automate the current ETL workflow.

6 Conclusion

Data in the agronomic domain are highly heterogeneous and dispersed. For agronomic researchers to make informed decisions in their daily work it is critical to integrate information at different scales. Current traditional information systems are not able to exploit such data (i.e., genes, proteins, metabolic

[1] Processing regular data update is a hard issue has the original databases do not always provide an automatic way to obtain the differential data between releases. From experience, we know that regularly rebuilding the entire knowledge base is for us a good alternative to avoid dealing with data diffs.

pathways, plant traits, and phenotypes), in efficient way. To this end, the application of Semantic Web, initiated in the biomedical domain, provides a good example to follow by capitalizing on previous experiences and addressing weaknesses.

To further build on this line of research in agronomy, we have developed AgroLD. We have demonstrated the advantages of AgroLD in data integration over multiple data sources using plant domain ontologies and Semantic Web technologies. To date, AgroLD contains 100M of triples created by transforming more than 50 datasets coming from 10 data and annotating with 10 ontologies. The impact of AgroLD is expected to grow with an increase in coverage (with respect to the species and the data sources) and user inputs. For instance, when user feedback and implementation of inference rules are put within a context that supports searching and recommendations, then we have the beginnings of a platform that can support automated hypotheses generation.

AgroLD is one of the first RDF linked open data knowledge-based system in the agronomic domain. It demonstrates a first step toward adopting the Semantic Web technologies to facilitate research by integrating numerous heterogeneous data and transforming them into explicitly knowledge thanks to ontologies. We expect AgroLD will facilitate the formulation of new scientific hypotheses to be validated with its knowledge-oriented approach.

Funding

This research was supported by the Computational Biology Institute of Montpellier (ANR-11-BINF-0002), the Institut Francais de Bioinformatique (ANR-11-INBS-0013), the Labex Agro (ANR-10-LABX-001-01) all bypass of the French ANR *Investissements d'Avenir* program.

Authors' contributions

AV designed and implemented the AgroLD project and wrote the manuscript. GT designed and implemented the API and the website. NEH contributed to the integration of data and set up of the RDF store. IC tested and formulated biological queries. VG contributed to the integration of data. CJ reviewed the manuscript. MR helped conceive the AgroLD project and reviewed the manuscript. PL conceived, designed, implemented the AgroLD project and wrote the manuscript. All the authors approved the final manuscript.

Acknowledgments

Authors thank the technical staffs of the South Green Bioinformatics platform for their support. Authors thank the providers of databases listed in Fig. 1, who kindly gave access to their publicly datasets. Authors thank the expert biologists and bioinformaticians who contributed to the testing sessions and helped us to improve the content of the system and the user interface. Authors specially thank Dr. Patrick Valduriez and Dr. Eric Rivals for their supports and advises in this project.

References

1. Goble C., and Stevens R. State of the nation in data integration for bioinformatics. J Biomed Inform. Elsevier; 2008;41: 687–693. doi:10.1016/j.jbi.2008.01.008

2. Harper L., Campbell J., Cannon E.K., Jung S., Main D., Poelchau M., Walls R., Andorf C., Arnaud E., Berardini T., Birkett C., Cannon S., Carson J., Condon B., Cooper L., Dunn N., Farmer A., Ficklin S., Grant D., et al. AgBioData Consortium Recommendations for Sustainable Genomics and Genetics Databases for Agriculture. Database. 2018; 1–7. (Submitted)

3. Ashburner M., Ball C.A., Blake J.A., Botstein D., Butler H., Cherry J.M., Davis A.P., Dolinski K., Dwight S.S., Eppig J.T., Harris M.A., Hill D.P., Issel-Tarver L., Kasarskis A., Lewis S., Matese J.C., Richardson J.E., Ringwald M., Rubin G.M., et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000;25: 25—29. doi:10.1038/75556

4. Cooper L., Walls R.L., Elser J., Gandolfo M.A., Stevenson D.W., Smith B., Preece J., Athreya B., Mungall C.J., Rensing S., Hiss M., Lang D., Reski R., Berardini T.Z., Li D., Huala E., Schaeffer M., Menda N., Arnaud E., et al. The plant ontology as a tool for comparative plant

409 anatomy and genomic analyses. Plant Cell Physiol. 2013;54: e1. doi:10.1093/pcp/pcs163

410 5. Shrestha R., Matteis L., Skofic M., Portugal A., McLaren G., Hyman G., and Arnaud E. Bridging
411 the phenotypic and genetic data useful for integrated breeding through a data annotation using
412 the Crop Ontology developed by the crop communities of practice. Front Physiol. 2012;3: 326.
413 doi:10.3389/fphys.2012.00326

414 6. Buttigieg P.L., Morrison N., Smith B., Mungall C.J., Lewis S.E., and ENVO Consortium. The
415 environment ontology: contextualising biological and biomedical entities. J Biomed Semantics.
416 2013;4: 43. doi:10.1186/2041-1480-4-43

417 7. Walls R.L., Deck J., Guralnick R., Baskauf S., Beaman R., Blum S., Bowers S., Buttigieg P.L.,
418 Davies N., Endresen D., Gandolfo M.A., Hanner R., Janning A., Krishtalka L., Matsunaga A.,
419 Midford P., Morrison N., Ó Tuama É., Schildhauer M., et al. Semantics in support of
420 biodiversity knowledge discovery: an introduction to the biological collections ontology and
421 related ontologies. PLoS One. 2014;9: e89606. doi:10.1371/journal.pone.0089606

422 8. Oellrich A., Walls R.L., Cannon E.K., Cannon S.B., Cooper L., Gardiner J., Gkoutos G. V,
423 Harper L., He M., Hoehndorf R., Jaiswal P., Kalberer S.R., Lloyd J.P., Meinke D., Menda N.,
424 Moore L., Nelson R.T., Pujar A., Lawrence C.J., et al. An ontology approach to comparative
425 phenomics in plants. Plant Methods. 2015;11: 10. doi:10.1186/s13007-015-0053-y

426 9. Wang Y., Wang Y., Wang J., Yuan Y., and Zhang Z. An ontology-based approach to integration
427 of hilly citrus production knowledge. Comput Electron Agric. Elsevier; 2015;113: 24–43.
428 doi:10.1016/J.COMPAG.2015.01.009

429 10. Lousteau-Cazalet C., Barakat A., Belaud J.-P., Buche P., Busset G., Charnomordic B., Dervaux
430 S., Destercke S., Dibie J., Sablayrolles C., and Vialle C. A decision support system for eco-
431 efficient biorefinery process comparison using a semantic approach. Comput Electron Agric.
432 Elsevier; 2016;127: 351–367. doi:10.1016/J.COMPAG.2016.06.020

433 11. Jonquet C., Toulet A., Arnaud E., Aubin S., Dzalé Yeumo E., Emonet V., Graybeal J., Laporte
434 M.A., Musen M.A., Pesce V., and Larmande P. AgroPortal: A vocabulary and ontology
435 repository for agronomy. Comput Electron Agric. 2018;144: 126–143.
436 doi:10.1016/j.compag.2017.10.012

437 12. Berners-lee T., Hendler J., and Lassila O. The Semantic Web. Sci Am. 2001;284: 35–43.

438 13. W3C. Resource Description Framework (RDF): Concepts and Abstract Syntax [Internet]. [cited
439 3 Apr 2010]. Available: http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/

440 14. W3C. RDF Schema 1.1 [Internet]. [cited 27 Apr 2018]. Available: https://www.w3.org/TR/rdf-
441 schema/

442 15. W3C. OWL 2 Web Ontology Language Structural Specification and Functional-Style Syntax
443 [Internet]. [cited 3 Apr 2010]. Available: http://www.w3.org/TR/2009/REC-owl2-syntax-
444 20091027/

445 16. The W3C SPARQL Working Group. SPARQL 1.1 Overview [Internet]. [cited 15 Apr 2013].
446 Available: http://www.w3.org/TR/sparql11-overview/

17. Luciano J.S., Andersson B., Batchelor C., Bodenreider O., Clark T., Denney C.K., Domarew C., Gambet T., Harland L., Jentzsch A., Kashyap V., Kos P., Kozlovsky J., Lebo T., Marshall S.M., McCusker J.P., McGuinness D.L., Ogbuji C., Pichler E., et al. The Translational Medicine Ontology and Knowledge Base: driving personalized medicine by bridging the gap between bench and bedside. J Biomed Semantics. 2011;2 Suppl 2: S1. doi:10.1186/2041-1480-2-S2-S1

18. Venkatesan A., Tripathi S., Sanz de Galdeano A., Blondé W., Lægreid A., Mironov V., and Kuiper M. Finding gene regulatory network candidates using the gene expression knowledge base. BMC Bioinformatics. 2014;15: 386. doi:10.1186/s12859-014-0386-y

19. Smith R.N., Aleksic J., Butano D., Carr A., Contrino S., Hu F., Lyne M., Lyne R., Kalderimis A., Rutherford K., Stepan R., Sullivan J., Wakeling M., Watkins X., and Micklem G. InterMine: a flexible data warehouse system for the integration and analysis of heterogeneous biological data. Bioinformatics. Oxford University Press; 2012;28: 3163–5. doi:10.1093/bioinformatics/bts577

20. Smith B., Ashburner M., Rosse C., Bard J., Bug W., Ceusters W., Goldberg L.J., Eilbeck K., Ireland A., Mungall C.J., Leontis N., Rocca-Serra P., Ruttenberg A., Sansone S.-A., Scheuermann R.H., Shah N., Whetzel P.L., Lewis S., and Lewis S. The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotechnol. Nature Publishing Group; 2007;25: 1251–1255. doi:10.1038/nbt1346

21. Noy N.F., Shah N.H., Whetzel P.L., Dai B., Dorf M., Griffith N., Jonquet C., Rubin D.L., Storey M.-A., Chute C.G., and Musen M.A. BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Res. 2009;37: W170-173. doi:10.1093/nar/gkp440

22. Belleau F., Nolin M.-A., Tourigny N., Rigault P., and Morissette J. Bio2RDF: towards a mashup to build bioinformatics knowledge systems. J Biomed Inform. Elsevier; 2008;41: 706–716. doi:10.1016/j.jbi.2008.03.004

23. Williams A.J., Harland L., Groth P., Pettifer S., Chichester C., Willighagen E.L., Evelo C.T., Blomberg N., Ecker G., Goble C., and Mons B. Open PHACTS: Semantic interoperability for drug discovery. Drug Discovery Today. 2012. pp. 1188–1198. doi:10.1016/j.drudis.2012.05.016

24. Momtchev V., Peychev D., Primov T., and Georgiev G. Expanding the Pathway and Interaction Knowledge in Linked Life Data. International Semantic Web Challenge. 2009.

25. Jupp S., Klein J., Schanstra J., and Stevens R. Developing a kidney and urinary pathway knowledge base. J Biomed Semantics. 2011;2 Suppl 2: S7. doi:10.1186/2041-1480-2-S2-S7

26. Jupp S., Malone J., Bolleman J., Brandizi M., Davies M., Garcia L., Gaulton A., Gehant S., Laibe C., Redaschi N., Wimalaratne S.M., Martin M., Le Novère N., Parkinson H., Birney E., and Jenkinson A.M. The EBI RDF platform: linked open data for the life sciences. Bioinformatics. 2014; 1–2. doi:10.1093/bioinformatics/btt765

27. Venkatesan A., El Hassouni N., Phillipe F., Pommier C., Quesneville H., Ruiz M., and Larmande P. Towards efficient data integration and knowledge management in the Agronomic domain. APIA'15: premiere Conference Applications Pratiques de l'Intelligence Artificielle. 2015.

486 28. Leonelli S., Davey R.P., Arnaud E., Parry G., and Bastow R. Data management and best practice
487 for plant science. Nat Publ Gr. Macmillan Publishers Limited; 2017;3: 1–4.
488 doi:10.1038/nplants.2017.86

489 29. Monaco M.K., Stein J., Naithani S., Wei S., Dharmawardhana P., Kumari S., Amarasinghe V.,
490 Youens-Clark K., Thomason J., Preece J., Pasternak S., Olson A., Jiao Y., Lu Z., Bolser D.,
491 Kerhornou A., Staines D., Walts B., Wu G., et al. Gramene 2013: Comparative plant genomics
492 resources. Nucleic Acids Res. 2014;42. doi:10.1093/nar/gkt1110

493 30. Magrane M., and Consortium U.P. UniProt Knowledgebase: A hub of integrated protein data.
494 Database. 2011;2011. doi:10.1093/database/bar009

495 31. Barrell D., Dimmer E., Huntley R.P., Binns D., O'Donovan C., and Apweiler R. The GOA
496 database in 2009 - An integrated Gene Ontology Annotation resource. Nucleic Acids Res.
497 2009;37. doi:10.1093/nar/gkn803

498 32. Hamelin C., Sempere G., Jouffe V., and Ruiz M. TropGeneDB, the multi-tropical crop
499 information system updated and extended. Nucleic Acids Res. 2013;41.
500 doi:10.1093/nar/gks1105

501 33. Droc G., Ruiz M., Larmande P., Pereira A., Piffanelli P., Morel J.B., Dievart A., Courtois B.,
502 Guiderdoni E., and Périn C. OryGenesDB: a database for rice reverse genetics. Nucleic Acids
503 Res. 2006;34: D736-40. doi:10.1093/nar/gkj012

504 34. Larmande P., Gay C., Lorieux M., Périn C., Bouniol M., Droc G., Sallaud C., Perez P., Barnola
505 I., Biderre-petit C., Martin J., Morel J.B., Johnson A.A.T., Bourgis F., Ghesquière A., Ruiz M.,
506 Courtois B., and Guiderdoni E. Oryza Tag Line, a phenotypic mutant database for the
507 Génoplante rice insertion line library. Nucleic Acids Res. 2008;36: 1022–1027.
508 doi:10.1093/nar/gkm762

509 35. Conte M.G., Gaillard S., Lanau N., Rouard M., and Périn C. GreenPhylDB: a database for plant
510 comparative genomics. Nucleic Acids Res. 2008;36: D991-998. doi:10.1093/nar/gkm934

511 36. Dereeper A., Homa F., Andres G., Sempere G., Sarah G., Hueber Y., Dufayard J.-F., and Ruiz
512 M. SNiPlay3: a web-based application for exploration and large scale analyses of genomic
513 variations. Nucleic Acids Res. 2015;43: W295-300. doi:10.1093/nar/gkv351

514 37. The Gene Ontology Consortium. Gene Annotation File (GAF) specification [Internet]. [cited 20
515 Mar 2018]. Available: http://geneontology.org/page/go-annotation-file-format-20

516 38. Sequence Ontology consortium. GFF3 Specification [Internet].

517 39. Gibbs R.A., Belmont J.W., Hardenbol P., Willis T.D., Yu F., Zhang H., Zeng C., Matsuda I.,
518 Fukushima Y., Macer D.R., Suda E., Stein L.D., Cunningham F., Kanani A., Thorisson G.A.,
519 Chakravarti A., Chen P.E., Cutler D.J., Kashuk C.S., et al. The International HapMap Project.
520 Nature. 2003;426: 789–796. doi:10.1038/nature02168

521 40. Danecek P., Auton A., Abecasis G., Albers C.A., Banks E., DePristo M.A., Handsaker R.E.,
522 Lunter G., Marth G.T., Sherry S.T., McVean G., and Durbin R. The variant call format and
523 VCFtools. Bioinformatics. 2011;27: 2156–8. doi:10.1093/bioinformatics/btr330

41. Juty N., Le Novère N., and Laibe C. Identifiers.org and MIRIAM Registry: community resources to provide persistent identification. Nucleic Acids Res. 2012;40: D580-6. doi:10.1093/nar/gkr1097

42. Smith B., Ceusters W., Klagges B., Kohler J., Kumar A., Lomax J., Mungall C., Neuhaus F., Rector A., and Rosse C. Relations in biomedical ontologies. Genome Biol. 2005;6: R46.

43. Cyganiak R. (National U. of I., and Bizer C. Pubby - A Linked Data Frontend for SPARQL Endpoints. 2008; Available: http://wifo5-03.informatik.uni-mannheim.de/pubby/

44. Heim P., Hellmann S., Lehmann J., Lohmann S., and Stegemann T. RelFinder: Revealing relationships in RDF knowledge bases. Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics). 2009. pp. 182–187. doi:10.1007/978-3-642-10543-2_21

45. Rietvelda L., and Hoekstraa R. The YASGUI Family of SPARQL Clients. Semant Web J. 2015;0: 1–10.

46. Elbedweihy K., Wrigley S.N., Ciravegna F., Reinhard D., and Bernstein A. Evaluating semantic search systems to identify future directions of research. The Semantic Web: ESWC 2012 Satellite Events. Springer; 2012. pp. 148–162.

47. Brooke J. SUS-A quick and dirty usability scale. Usability Eval Ind. London; 1996;189: 4–7.

48. Wilkinson M.D., Dumontier M., Aalbersberg Ij.J., Appleton G., Axton M., Baak A., Blomberg N., Boiten J.-W., da Silva Santos L.B., Bourne P.E., Bouwman J., Brookes A.J., Clark T., Crosas M., Dillo I., Dumon O., Edmunds S., Evelo C.T., Finkers R., et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci Data. 2016;3. doi:10.1038/sdata.2016.18

49. Venkatesan A., Kim J.-H., Talo F., Ide-Smith M., Gobeill J., Carter J., Batista-Navarro R., Ananiadou S., Ruch P., McEntyre J., Venkatesan A., Kim J.-H., Talo F., Ide-Smith M., Gobeill J., Carter J., Batista-Navarro R., Ananiadou S., Ruch P., et al. SciLite: a platform for displaying text-mined annotations as a means to link research articles with biological data. Wellcome Open Res. 2016;1: 25. doi:10.12688/wellcomeopenres.10210.1

50. IC4R Project Consortium, Hao L., Zhang H., Zhang Z., Hu S., and Xue Y. Information Commons for Rice (IC4R). Nucleic Acids Res. 2016;44: D1172–D1180. doi:10.1093/nar/gkv1141

51. Petryszak R., Keays M., Tang Y.A., Fonseca N.A., Barrera E., Burdett T., Füllgrabe A., Fuentes A.M.P., Jupp S., Koskinen S., Mannion O., Huerta L., Megy K., Snow C., Williams E., Barzine M., Hastings E., Weisser H., Wright J., et al. Expression Atlas update - An integrated database of gene and protein expression in humans, animals and plants. Nucleic Acids Res. 2016;44: D746–D752. doi:10.1093/nar/gkv1045

52. Lee T., Oh T., Yang S., Shin J., Hwang S., Kim C.Y., Kim H., Shim H., Shim J.E., Ronald P.C., and Lee I. RiceNet v2: An improved network prioritization server for rice genes. Nucleic Acids Res. 2015;43: W122–W127. doi:10.1093/nar/gkv253