AgroLD: une base de connaissances pour l'étude du phénome des plantes cultivées

Pierre Larmande, Tagny Gildas, Manuel Ruiz

AgroLD: une base de connaissances pour l'étude du phénome des plantes cultivées

Pierre Larmande[*,***], Gildas Tagny Ngompe[**,***] Manuel Ruiz[**,***]

[*]UMR DIADE, IRD, Univ. Montpellier, 911 av Agropolis, 34398 Montpellier, France
pierre.larmande@ird.fr,
[**]UMR AGAP, CIRAD,INRAE, Univ. Montpellier,
Avenue Agropolis, 34398 Montpellier Cedex 5 Montpellier, France
manuel.ruiz@cirad.fr, gildas.tagny_ngompe@cirad.fr
[***]SOUTH GREEN BIOINFORMATICS PLATFORM, av Agropolis, 34398 Montpellier, France

Résumé. Les récents progrès des technologies à haut débit ont entraîné une explosion de la quantité de données dans le domaine agronomique. Il est urgent d'intégrer efficacement des informations complémentaires pour comprendre le système biologique dans sa globalité. Nous avons développé AgroLD, une base de connaissances qui exploite la technologie du Web sémantique et des ontologies du domaine biologique pertinentes, pour intégrer les informations sur les espèces végétales et faciliter ainsi la formulation de nouvelles hypothèses scientifiques. Nous présentons des résultats sur le processus d'intégration et sur la plateforme visualisation des données, qui était initialement axé sur la génomique, la protéomique et la phénomique.

1 Introduction

La compréhension des interactions génotype-phénotype est un des axes les plus importants de la recherche en agronomie dont l'un des objectifs est d'accélérer la reproduction des caractères importants pour la production agricole. Or ces interactions sont complexes à identifier car elles s'expriment à différentes échelles moléculaires dans la plante et subissent de fortes influences de la part des facteurs environnementaux. Les technologies d'analyse haut-débit ne permettent de capturer que partiellement cette dynamique. Même si ces technologies sont de plus en plus performantes dans l'acquisition de données, notre connaissance du système reste encore parcellaire pour pouvoir comprendre les relations complexes existant entre les différents éléments moléculaires responsables de l'expression du phénome -ensemble des phénotypes observés pour un individu- . Cet objectif ne peut être atteint qu'en intégrant des informations de différents niveaux dans un modèle intégrateur utilisant une approche systémique afin de comprendre le fonctionnement réel d'un système biologique. Aujourd'hui, le Web sémantique propose des technologies pour l'intégration de données hétérogènes et leur transformation en connaissances explicites grâce aux ontologies. Notre hypothèse repose sur l'idée que proposer des graphes de connaissances fondées sur les données et informations produites permettrait de formuler plus aisément des hypothèses de recherche permettant de lier le

génotype au phénotype. En prenant le riz comme espèce modèle principale, l'objectif sera de construire des réseaux d'interactions moléculaires à partir de données éparses afin d'identifier les gènes clés pour l'amélioration des plantes. Diverses approches seront mises à contribution relatives à l'intégration de données, à l'enrichissement des connaissances et à des applications sur les graphes de connaissances.

2 La plateforme AgroLD

2.1 Contexte

Nous avons développé AgroLD [1] (Venkatesan et al., 2018), une base de connaissances reposant sur les technologies du Web sémantique et exploitant des ontologies du domaine biologique, afin d'intégrer des données issues de plusieurs espèces de plantes présentant un intérêt important pour la communauté scientifique, comme par exemple le riz, le blé et arabidopsis. De telles initiatives équivalentes existaient dans le domaine biomédical et bioinformatique, citons Bio2RDF (Belleau et al., 2008), EBI RDF (Jupp et al., 2014), ou encore Uniprot RDF (Redaschi et the UniProt Consortium, 2009), mais aucune dans le domaine agronomique. Nous présentons aujourd'hui, les résultats du projet, qui portait initialement sur la génomique, la protéomique et la phénomique. AgroLD contient plus de 100 millions de triplets créée à partir de plus de 50 jeux de données provenant d'une dizaine de sources de données. Pour cette phase, chaque jeu de données a été transformé à partir de sources sélectionnées et annotées sémantiquement en réutilisant les champs textuels correspondant avec des termes d'ontologies lorsqu'ils ont été fournis par la source d'origine.

L'objectif d'AgroLD est d'offrir une plate-forme de connaissances spécifiques du domaine agronomique afin de répondre à des questions biologiques complexes. De telles questions peuvent concerner le rôle de gènes spécifiques dans les mécanismes de résistance aux maladies des plantes ou de caractères de production identifiés à partir des analyses GWAS [2]. Afin de rendre AgroLD accessible par un plus grand nombre d'utilisateurs, nous avons également développé une application Web proposant plusieurs interfaces de requêtes. Tout d'abord une interface simple qui permet aux utilisateurs d'effectuer des rechercher par mots-clés sur l'ensemble des valeurs de la base et ainsi de parcourir le contenu de la base de connaissance. Puis une interface de recherche avancée qui permet de combiner du texte libre et des filtres base sur les types de classes et propriétés ainsi que des services Web externes proposant ainsi une interface d'agrégation de données distribuées. AgroLD possède également une interface de visualisation des graphes qu'il est possible de configurer pour mettre en valeur certains types de relations. Finalement, un éditeur SPARQL propose un environnement interactif pour formuler des requêtes et manipuler des résultats.

1. Agronomic Linked Data - `http://www.agrold.org`
2. GWAS, Genome Wide Association Studies, sont des expérimentations biologiques impliquant des méthodes statistiques permettant de corréler un caractère phénotypique a une ou plusieurs région du génome.

2.2 Inventaire des sources de données intégrées

Le cadre conceptuel de la connaissance est basé sur des ontologies bien établies dans le domaine telles que Gene Ontology (Ashburner et al., 2000), une ontologie sur la fonction des gènes ; Plant Ontology (Plant et Consortium, 2002), une ontologie sur l'anatomie des plantes ; Plant Trait Ontology (Cooper et al., 2018), une ontologie sur les caractères phénotypiques des plantes. La majorité de ces ontologies sont hébergées par le projet OBO Foundry (Smith et al., 2007). En outre, compte tenu de la portée de l'effort, nous avons décidé de construire AgroLD en plusieurs phases. La phase actuelle (première phase) couvre les informations sur les gènes, les protéines, les prédictions de gènes homologues, les voies métaboliques, des phénotypes de plantes et le matériel génétique. A ce stade nous avons intégré des données issues de plusieurs ressources telles que Gramene, qui identifie les gènes chez les plantes cultivées ; UniProt, qui répertorie les protéines et leurs fonctions chez tous les êtres vivants ; Gene Ontology Annotation qui identifie les associations de concepts de Gene Ontology avec des gènes ou des protéines. Le choix de ces sources a été guidé par la communauté biologique avec qui nous collaborons. Elles sont en effet très utilisées et bénéficient d'un fort impact sur la confiance des données. Nous avons également intégré des ressources développées par la plateforme montpelliéraine SouthGreen [3]. Ces ressources regroupent des données expérimentales produites par les chercheurs montpelliérains et leurs partenaires. Le tableau 1 donne un aperçu des espèces et sources intégrées.

2.3 Contributions dans le domaine de l'ingénierie des connaissances

2.3.1 Vers une automatisation des transformations RDF

Nos contributions portent sur la création de différents pipelines de transformation RDF pour des grands jeux de données agronomiques. Même si de nombreux outils étaient disponibles au sein de la communauté du Web Sémantique, parmi eux citons datalift [4] ou des implementation de csv2rdf [5] ou encore RML.io [6]. Aucun n'était adapté pour prendre en compte la complexité des formats de fichiers du domaine biologique (par exemple le format VCF) ou même la complexité des informations qu'ils pouvaient contenir. Un exemple très simple illustre cette complexité à travers le format GFF (Generic Feature Format) [7] qui représente les données génomique dans un format de type TSV. Il contient une colonne ayant des informations de type *clé=valeur*, de longueur variables et différentes selon les sources de données. Dans ce cas, il est nécessaire d'adapter la transformation en fonction de la source de données. Par ailleurs, le volume important des sources de données était un facteur limitant des outils sus-mentionnés.

3. South Green Bioinformatic Platform - `http://southgreen.fr/`
4. `https://project.inria.fr/datalift`
5. `https://www.w3.org/TR/csv2rdf/`
6. `https://rml.io/`
7. `http://gmod.org/wiki/GFF3`

Sources de données	**Format de fichier**	**Nb Tuples**	**Espèces**	**Ontologies utilisées**	**Nb de triplets produits**
Oryzabase	TSV	17K	R	GO,PO,TO	153 K
GOA	GAF	1, 160K	All	GO	2, 700 K
OryGenes	GFF	1, 100K	R, S, A,	GO, SO	5, 172 K
Gramene	TSV	1, 718K	All	All	30, 000 K
Uniprot	TSV	1, 400K	All	GO, PO	50, 000 K
OryzaTagLine	TSV	22K	R	PO, TO, CO	300 K
TropGene	TSV	2K	R	PO, TO, CO	20 K
GreenPhyl	TSV	100K	R,A	GO, PO	700 K
SNiPlay	VCF	16K	R	GO	16,00 0K
Q-TARO	TSV	2K	R	PO, TO	20 K
TOTAL					**105,065K**

TAB. 1 – ***Les espèces et les sources de données intégrées dans AgroLD.*** *Le nombre de tuples donne une idée du nombre d'éléments que nous avons annotés à partir des sources de données (par exemple, 1, 160K Gene Ontology annotations). Especes et Ontologies sont référencées suivant R = riz, W = blé, A = Arabidopsis, S = sorgho, M = maïs, GO = gene ontology, PO = plant ontology, TO = plant trait ontology, EO = plant environnment ontology, SO = sequence ontology, CO = crop ontology (caractères spécifiques des plantes). 134 ontologies)*

Dans ce contexte, nous avons développé des modèles de transformation RDF adaptés à une plus large palette de standards de données en génomique et phénomique tels que GFF, GAF [8], VCF [9] et travaillons actuellement à packager ces modèles dans une API [10]. Ces standards représentent une première étape, car ils sont en effet, les plus utilisés dans la communauté. Nous comptons développer de nouveaux modèles pour d'autres standards de données, notamment pour les données phénotypiques MIAPPE [11].

2.3.2 Annotation sémantique des données avec des bio-ontologies

Pour cette phase de transformation, chaque jeu de données a été téléchargé à partir de sources sélectionnées et annoté sémantiquement avec des URI de termes ontologiques en réutilisant les identifiants d'ontologie lorsqu'ils ont été fournis par la source d'origine. De plus, lorsque cela était possible, nous avons utilisé des annotations sémantiques déjà présentes dans les jeux de données, telles que, par exemple, des gènes ou des traits annotés respectivement avec des identifiants Gene Ontology [12] ou Trait Ontology. Dans ce cas, nous avons généré des

8. le Gene Ontology Annotation File `http://geneontology.org/page/go-annotation-file-format-20`
9. le Variant Call Format`https://samtools.github.io/hts-specs/VCFv4.2.pdf`
10. `https://github.com/SouthGreenPlatform/AgroLD_ETL`
11. MIAPPE : Minimum Information About a Plant Phenotyping Experiment - `https://www.miappe.org/`
12. GO :0005524 est transformé en URI :`http://purl.obolibrary.org/obo/GO_0005524`

propriétés supplémentaires avec les ontologies correspondantes, ajoutant ainsi 22% de triplets supplémentaires validés manuellement (voir les détails dans le tableau 1). Les versions OWL des ontologies candidates ont été directement chargées dans la base de connaissances, mais leurs triplets ne sont pas comptés dans le total.

Par ailleurs, nous avons utilisé l'API de service Web AgroPortal (Jonquet et al., 2018), pour enrichir les données en annotations sémantiques. Par exemple, pour identifier des concepts ontologiques dans les données comme l'organe d'une plante [13]) ou un caractère phénotypique [14]). Pour cela, nous avons développé, *Table2Annotation* (Larmande et Jibril, 2020), une application spécifique pour traiter les formats de fichiers semi-structurés (tsv, csv, excel), mieux contrôler l'annotation sémantique faite par AgroPortal et y gérer les différents cas particuliers d'annotations pour un résultat optimal.

2.3.3 Méthodes de liage d'entités issus de graphes distincts

Les graphes RDF partagent un espace de noms commun [15] et sont nommés d'après les sources de données correspondantes. Les entités dans les graphes RDF sont liées par des bases d'URI communes. En général, nous avons construit les URI en nous référant à Identifiers.org, qui fournit des patrons de conception pour chaque source enregistrée. Par exemple, les gènes intégrés à partir de la source Ensembl Plant sont identifiés par l'URI de base [16]. Lorsqu'elles ne sont pas fournies par Identifiers.org, de nouvelles URI sont construites ; dans ce cas, les URI prennent la forme [17]. Par ailleurs, les propriétés reliant les entités sont construites sous la forme [18].
Afin de lier des entités similaires issues de sources différentes, nous avons utilisé l'approche basée sur *l'identification de la clé* qui est la plus courante. Son principe est d'analyser les URI afin de rechercher des motifs similaires dans la partie terminale de l'URI. De plus, nous avons également respecté *l'approche URI commune*, qui recommande d'utiliser le même patron d'URI pour deux entités similaires. De ce fait, pour une même entité, cela nous a permis d'agréger des informations issues de différents graphes RDF. Par ailleurs, nous avons utilisé des liens de références croisées en les transformant en URI et en reliant la ressource au prédicat *rdfs :seeAlso*. Cela augmente considérablement le nombre de liens sortants, rendant AgroLD mieux intégrée avec d'autres sources de données. À l'avenir, nous comptons mettre en œuvre une approche basée sur la similarité des propriétés pour identifier les correspondances entre les entités ayant des URI différents.

Afin de faire correspondre les différents types de données et propriétés, nous avons développé un schéma [19] qui associe les classes et propriétés identifiées dans AgroLD avec des onto-

13. *leaf* est annoté avec le concept PO_0025034 avec l'URI :`http://purl.obolibrary.org/obo/PO\_0025034`
14. *plant height* serait annoté avec le concept ayant TO_0000207 pour URI :`http://purl.obolibrary.org/obo/TO\_0000207`
15. `http://www.southgreen.fr/agrold/`
16. `http://identifiers.org/ensembl.plant/`
17. `http://www.southgreen.fr/agrold/[resourceNamespace]/[identifier]`
18. `http://www.southgreen.fr/agrold/vocabulary/[property]`
19. `https://github.com/SouthGreenPlatform/AgroLD_ETL/tree/master/model`

logies correspondantes. Par exemple, la classe Protein [20] est associée à la classe polypeptide [21] de SO avec la propriété *owl : equivalentClass*. Des mappings similaires ont été réalisés pour les propriétés, par exemple, les classes *Protein* et *Gene* sont liées aux classes de l'ontologie *molecular function* de GO par la propriété *has_function* [22], avec comme propriété *owl : equivalentProperty*. Lorsqu'une propriété équivalente n'existait pas, nous l'avons associée avec la propriété de niveau supérieur avec *rdfs : subPropertyOf*. Par exemple, la propriété *has_trait* [23], relie les entités aux termes TO équivalent. Elle est associée à une propriété plus générique. Au total, 55 mappings ont été identifiés.

2.4 Faciliter l'accès aux données liées

En matière d'accès aux graphes de données, même si le langage SPARQL est efficace pour construire les requêtes, il reste difficile à prendre en main pour nos utilisateurs principaux (bioinformaticiens et biologistes). Ainsi, nous avons proposé un modèle d'architecture implémentant divers éléments constituant de systèmes de recherche sémantique (i.e., formulation de requêtes basé sur des patrons, visualisation sous forme de graphe, outils de recherche d'information). Ainsi la plateforme AgroLD fournit 4 points d'entrée :

— **Quick Search** [24], un plugin de recherche à facette mis à disposition par Virtuoso, qui permet aux utilisateurs d'effectuer des recherches par mots-clés et de parcourir le contenu d'AgroLD en naviguant dans les liens ;

— **SPARQL Editor** [25], un éditeur de requêtes SPARQL qui fournit un environnement interactif pour la formulation de requêtes SPARQL. Nous avons développé l'éditeur en se basant sur les outils YASQE et YASR (Rietveld et Hoekstra, 2015) et l'avons adapté pour notre système. Par ailleurs, nous avons proposé une liste de patrons de requêtes modulaires et personnalisables en fonction des besoins des utilisateurs qui peuvent être automatiquement exécutées à travers l'éditeur ;

— **Explore Relationships** [26], est une version modifiée de RelFinder (Heim et al., 2009), qui permet aux utilisateurs d'explorer et de visualiser les relations existantes entre entités ;

— **Advanced Search** [27], un formulaire proposant des recherches spécifiques par entité et possédant un moteur d'agrégation de ressources externes. Le formulaire Advanced Search est basé sur une API REST [28]. Le but de ce formulaire est de fournir aux biologistes un outil permettant d'interroger la base de connaissances tout en masquant les

20. http://www.southgreen.fr/agrold/resource/Protein
21. http://purl.obolibrary.org/obo/SO_0000104
22. http://www.southgreen.fr/agrold/vocabulary/has_function
23. http://www.southgreen.fr/agrold/vocabulary/has_trait
24. http://www.agrold.org/quicksearch.jsp
25. http://www.agrold.org/sparqleditor.jsp
26. http://www.agrold.org/relfinder.jsp
27. http://www.agrold.org/advancedSearch.jsp
28. http://www.agrold.org/api-doc.jsp

aspects techniques de la formulation de requêtes SPARQL. L'intérêt de coupler API et formulaire est de pouvoir combiner de manière interactive des recherches dans la base de connaissances et dans des services externes à la fois par l'interface utilisateur mais également par la programmation.

3 Conclusion

Actuellement, de nouveaux jeux de données sont en cours d'intégration. Ils portent sur les réseaux d'interaction protéine-protéine, les facteurs de transcription et réseaux de co-expression afin d'étendre les connaissances sur les mécanismes moléculaires. De nombreux développements sont également réalisés au niveau des interfaces de requêtes, notamment au niveau de la visualisation des graphes afin de fournir des outils plus dynamiques, interactifs et contextualisés. Enfin, une attention particulière est portée sur la qualité des données intégrées. Des méthodes de liage et de machine learning sont développées pour rechercher des liens et des ressources similaires dans la base de connaissances ou dans des ressources externes.

Références

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin, et G. Sherlock (2000). Gene ontology : tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet 25*(1), 25–29.

Belleau, F., M.-A. Nolin, N. Tourigny, P. Rigault, et J. Morissette (2008). Bio2RDF : towards a mashup to build bioinformatics knowledge systems. *Journal of biomedical informatics 41*(5), 706–16.

Cooper, L., A. Meier, M. A. Laporte, J. L. Elser, C. Mungall, B. T. Sinn, D. Cavaliere, S. Carbon, N. A. Dunn, B. Smith, B. Qu, J. Preece, E. Zhang, S. Todorovic, G. Gkoutos, J. H. Doonan, D. W. Stevenson, E. Arnaud, et P. Jaiswal (2018). The Planteome database : An integrated resource for reference ontologies, plant genomics and phenomics. *Nucleic Acids Research 46*(D1).

Heim, P., S. Hellmann, J. Lehmann, S. Lohmann, et T. Stegemann (2009). RelFinder : Revealing relationships in RDF knowledge bases. In *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, Volume 5887 LNCS, pp. 182–187. Citation Key : Heim2009 ISSN : 03029743.

Jonquet, C., A. Toulet, E. Arnaud, S. Aubin, E. Dzalé Yeumo, V. Emonet, J. Graybeal, M. A. Laporte, M. A. Musen, V. Pesce, et P. Larmande (2018). AgroPortal : A vocabulary and ontology repository for agronomy. *Computers and Electronics in Agriculture 144*(October 2016), 126–143.

Jupp, S., J. Malone, J. Bolleman, M. Brandizi, M. Davies, L. Garcia, A. Gaulton, S. Gehant, C. Laibe, N. Redaschi, S. M. Wimalaratne, M. Martin, N. Le Novère, H. Parkinson, E. Bir-

ney, et A. M. Jenkinson (2014). The EBI RDF platform : linked open data for the life sciences. *Bioinformatics (Oxford, England)*, 1–2.

Larmande, P. et K. M. Jibril (2020). Enabling Fast Annotation Process With Table2Annotation Tool. *bioRxiv*, 2020.04.03.023069.

Plant, T. et O. Consortium (2002). The Plant Ontology Consortium and plant ontologies. *Comparative and functional genomics 3*(2), 137–42. Citation Key : Plant2002.

Redaschi, N. et the UniProt Consortium (2009). Uniprot in RDF : Tackling data integration and distributed annotation with the semantic web. *Nature Prec*.

Rietveld, L. et R. Hoekstra (2015). The YASGUI Family of SPARQL Clients. *Semantic Web Journal*. Citation Key : Rietveld2015YASGUI.

Smith, B., M. Ashburner, C. Rosse, J. Bard, W. Bug, W. Ceusters, L. J. Goldberg, K. Eilbeck, A. Ireland, C. J. Mungall, N. Leontis, P. Rocca-Serra, A. Ruttenberg, S.-A. Sansone, R. H. Scheuermann, N. Shah, P. L. Whetzel, et S. Lewis (2007). The OBO Foundry : coordinated evolution of ontologies to support biomedical data integration. *Nat Biotech 25*(11), 1251–1255.

Venkatesan, A., G. Tagny Ngompe, N. E. Hassouni, I. Chentli, V. Guignon, C. Jonquet, M. Ruiz, et P. Larmande (2018). Agronomic Linked Data (AgroLD) : A knowledge-based system to enable integrative biology in agronomy. *PLOS ONE 13*(11), 1–17.

Summary

Recent advances in high-throughput technologies have resulted in tremendous increase in the amount of data in the agronomic domain. There is an urgent need to effectively integrate complementary information to understand the biological system in its entirety. We have developed AgroLD, a knowledge graph that exploits the Semantic Web technology and some of the relevant standard domain ontologies, to integrate information on plant species and in this way facilitating the formulation of new scientific hypotheses. We present some integration results of the project, which initially focused on genomics, proteomics and phenomics.