Intégration de données multi-échelles et extraction de connaissances en agronomie : exemples et perspectives

Pierre Larmande

HABILITATION Á DIRIGER DES RECHERCHES

UNIV. MONTPELLIER, ECOLE DOCTORALE I2S, INFORMATIQUE

Intégration de Données Multi-Échelles et Extraction de Connaissances en Agronomie : Exemples et Perspectives

Pierre Larmande
(ORCID :0000-0002-2923-9790)
pierre.larmande@ird.fr

IRD
UMR DIADE
Montpellier, France

Version déposée, en attente de validation des rapporteurs

JURY

Juliette DIBIE-BARTHELEMY	Professeur, INRAE, Paris	Rapporteur
Catherine FARON-ZUCKER	Maître de Conférence, HDR, Univ. de Sophia Antipolis, Nice	Rapporteur
Claire NEDELLEC	Directrice de Recherche, INRAE, Paris	Rapporteur
Isabelle MOUGENOT	Maître de Conférence, HDR, Univ. de Montpellier	Examinateur
Manuel RUIZ	Directeur de Recherche, CIRAD, Montpellier	Examinateur
Hadi QUESNEVILLE	Directeur de Recherche, INRAE, Versailles	Examinateur
Thérèse LIBOUREL	Professeur, Univ. de Montpellier	Marraine scientifique

UNIV. MONTPELLIER, ECOLE DOCTORALE I2S, INFORMATIQUE

Résumé

Montpellier, France
UMR DIADE

Habilitation á Diriger des Recherches

Intégration de Données Multi-Échelles et Extraction de Connaissances en Agronomie : Exemples et Perspectives

by Pierre LARMANDE

La compréhension des relations génotype-phénotype est un des axes les plus importants de la recherche en agronomie. Les nouveaux défis consistent à comprendre ces relations existant entre les différents éléments moléculaires responsables de l'expression du phénome. Ceux-ci, nous semble t'il, ne peuvent être relevés qu'en intégrant des informations de différents niveaux dans un modèle global utilisant une approche systémique afin de comprendre le fonctionnement réel d'un système biologique. Les technologies d'analyses haut-débit récentes ne permettent de capturer que partiellement cette dynamique. Même si ces technologies permettent d'aller toujours plus loin dans l'obtention de nouvelles données, une réflexion sur la centralisation de celles-ci au sein de supports dédiés et sur la standardisation des formats devrait permettre de les regrouper efficacement, et contribuer de ce fait à l'amélioration des connaissances. En effet, il s'avère que celles-ci restent encore parcellaires pour élucider les mécanismes moléculaires qui régissent l'expression de caractères phénotypiques complexes.

Mon projet de recherche s'inscrit dans cette ligne de réflexions et aborde le problème suivant : Comment structurer et gérer la complexité des données biologiques afin d'en extraire de la connaissance permettant d'identifier les mécanismes moléculaires contrôlant l'expression de phénotypes chez les plantes.

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 modèle, 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 la conception de graphes de connaissances.

Dans ce processus, une première voie consistera à transformer et intégrer dynamiquement les données pertinentes au sein d'une base de connaissance pour les rendre plus facilement utilisables dans les traitements analytiques (en terme algorithmique). Une deuxième voie consistera à proposer de nouvelles méthodes d'enrichissement des connaissances. Dans un premier temps, en mobilisant des méthodes d'annotation sémantique ; puis, en développant de nouvelles méthodes de liage de données afin d'exploiter de nouveaux liens entre les différents graphes générés et ainsi produire un réseau d'interactions qui permettra la découverte de nouvelles connaissances. Enfin, afin de permettre une recherche d'information efficace, plusieurs méthodes et algorithmes de priorisation de gènes candidats seront évaluées et proposées au sein des graphes de connaissances disponibles.

Remerciements

Tout d'abord, je voudrai remercier mes collègues de l'UMR DIADE, de la plate-forme South-Green et du LIRMM pour leurs discussions et échanges fructueux.

Je remercie les membres de l'équipe RICE et de l'équipe FADO pour leur influence positive sur mes recherches et sur la construction de mon projet de recherche.

J'ai une pensée particulière pour mon directeur d'unité, Alain Ghesquière, qui a su me faire confiance et m'encourager durant ces 10 dernières années.

Au cours de ces années, j'ai également eu l'opportunité et le privilège de collaborer avec de nombreux chercheurs Français et étrangers. Ces collaborations m'ont beaucoup appris et m'ont aidé à construire ce projet.

Je remercie également chaleureusement Thérèse Libourel, qui m'a encadré durant ma thèse et plus récemment m'a soutenu dans ce projet d'écriture.

Je voudrai également remercier les membres du jury d'avoir accepté d'évaluer mon HDR.

Enfin et surtout, je remercie ma famille pour leur soutien sans faille et leur amour.

Table des matières

Résumé **iii**

Remerciements **v**

I Données et Connaissances en Agronomie : Recherches effectuées **1**

1 Préambule **3**
1.1 Organisation du mémoire . . . 3
1.2 Déroulement de carrière . . . 4
1.3 Soutiens financiers et personnes impliquées . . . 6
1.4 Collaborations . . . 7

2 Introduction **9**

3 Contexte scientifique, Enjeux et Problématique **13**
3.1 Biologie moléculaire et génétique . . . 13
3.1.1 Caractérisation des relations génotype-phénotype . . . 14
3.1.2 Les mécanismes qui régulent l'expression des gènes . . . 15
3.1.3 La biologie moléculaire dans le cas du riz . . . 17
3.2 Les enjeux computationnels . . . 18
3.2.1 La révolution des technologies haut-débit . . . 18
3.2.2 Analyses de variabilité . . . 19
3.3 L'intégration de données . . . 20
3.3.1 L'hétérogénéité des systèmes . . . 20
3.3.2 L'évolution des approches d'intégration de données . . . 21
3.4 Représentation des données . . . 24
3.4.1 Rappel sur le web de données . . . 24
3.4.2 Exemples de représentation des données en biologie . . . 25
3.5 Extraction de connaissances biologiques . . . 28
3.5.1 Méthodes d'extraction d'entités nommées . . . 28
3.5.2 Utilisation de méthodes d'apprentissage profond pour l'extraction d'entités nommées . . . 29
3.5.3 Méthodes d'extraction de relations . . . 30

4 Synthèse des activités de recherche et résultats obtenus **33**
4.1 Propositions d'approches décentralisées pour l'interopérabilité des bases de données agronomiques . . . 33
4.1.1 Une architecture de médiation est elle adaptée dans ce contexte . . . 34
4.1.2 Comment la composition de services Web peut faciliter l'intégration . . . 35
4.1.3 Conclusion . . . 36
4.2 Enrichissement automatique de mappings BD-RDF pour faciliter la réécriture de requêtes . . . 37
4.2.1 État de l'art des approches de mappings BD-RDF . . . 38

4.2.2 Enrichissement sémantique des mappings BDR-RDF . . . 39
4.2.3 Méthode de réécriture de requêtes SPARQL à partir de contraintes de schémas 42
4.2.4 Évaluation d'approches de mapping NoSQL-RDF . . . 43
4.2.5 Conclusion . . . 44
4.3 Passage à l'échelle dans la gestion des données génomiques . . . 44
4.3.1 Comment combiner le stockage massif et les performances de requêtes . . . 45
4.3.2 Conclusion . . . 46
4.4 Vers de nouvelles approches d'intégration sémantique des données agronomiques . 46
4.4.1 La plateforme AgroPortal . . . 47
4.4.2 La plateforme AgroLD . . . 48
4.4.3 Conclusion . . . 56

II Projet et perspectives **59**

5 Projet **61**
5.1 Intégration de données et annotation sémantique . . . 61
5.1.1 Intégration dynamique des données . . . 61
5.1.2 Annotation sémantique . . . 62
5.2 Extraction et exploitation de la connaissance . . . 63
5.2.1 Extraction d'entités biologiques et de relations . . . 63
5.2.2 Liage des données . . . 67
5.2.3 Raisonnement sur les données . . . 73
5.3 Applications sur les graphes de connaissances . . . 74
5.3.1 Priorisation de gènes candidats . . . 74
5.3.2 Analyse fonctionnelle des réseaux d'interactions moléculaires . . . 75

6 Conclusion et perspectives **77**
6.1 Conclusion . . . 77
6.2 Perspectives . . . 78

III Annexes **81**

7 Curriculum Vitae **83**
7.1 Identité . . . 83
7.2 Formation . . . 83
7.3 Expérience professionnelle . . . 84
7.4 Éléments marquants du CV . . . 84
7.5 Projets scientifiques . . . 85
7.6 Prototypage . . . 86
7.7 Animation de la recherche . . . 87
7.8 Activités d'enseignement . . . 89
7.9 Encadrements . . . 89

8 Liste des publications **93**

Bibliographie **99**

IV Sélection de publications **119**

A Sélection de publications **121**
A.1 Journal 121

Liste des abréviations

ADN	Acide Désoxyribo Nucléique
ARN	Acide Ribo Nucléique
API	Application Programming Interface
BDR	Base de Données Relationnelles
CAAS	Chinese Academy of Agricultural Science
CGIAR	Consultative Group on International Agricultural Research
CNV	Copy Number Variations
CRF	Conditional Random Fields
ETL	Extraction Transform Load
GCP	Generation Challenge Program
GFVO	Genomic Feature and Variation Ontology
GAF	Gene Ontology Annotation File
GAV	Global As View
GFF	Generic Feature Format
GVCF	Genomic VCF
HTTP	HyperText Transfer Protocol
IRI	International Resource Identifier
IRGSP	International Rice Genome Sequencing Project
IRRI	International Rice Research Institute
LAV	Local As View
LSTM	Long Short Term Memory
LOV	Linked Open Vocabulary
MIAPPE	Minimum Information About a Plant Phenotyping Experiment
NER	Named Entity Recognition
NGS	Next Generation Sequencing
OBO	Open Biomedical Ontologies
QTL	Quantitative Trait Loci
REST	Representational state transfer
RDF	Resource Description Framework
SGBD	Système de Gestion de Base de Données
SNP	Single Nucleotide Polymorphism
SQR	Systèmes de Questions-Réponses
URI	Uniform Resource Identifier
VCF	Variant Call Format
XML	eXtensible Markup Language

A Sophie, Nina et Salomé qui m'ont soutenu et encouragé depuis de nombreuses années . . .

Première partie

Données et Connaissances en Agronomie : Recherches effectuées

Chapitre 1

Préambule

Ce mémoire va se consacrer à la présentation d'un projet de recherche qui a progressivement mûri au cours de plusieurs années d'activités diverses et de recherches effectuées dans les domaines de l'intégration de données et de l'extraction de connaissances biologiques appliqués à l'agronomie. Depuis mon doctorat (soutenu en 2007) « Mutualiser et partager, un défi pour la génomique fonctionnelle végétale », le dénominateur commun de mes recherches peut s'exprimer en « **comment extraire, intégrer et formaliser la connaissance biologique** » nécessaire à une meilleure compréhension du diptyque génotype-phénotype.

Derrière ce cadre, nombre de questions, nombre d'écueils, de verrous, nombre de propositions se sont faits jour et ont contribué au projet ; il seront dans la mesure du possible, bien que parfois brièvement, relatés.

1.1 Organisation du mémoire

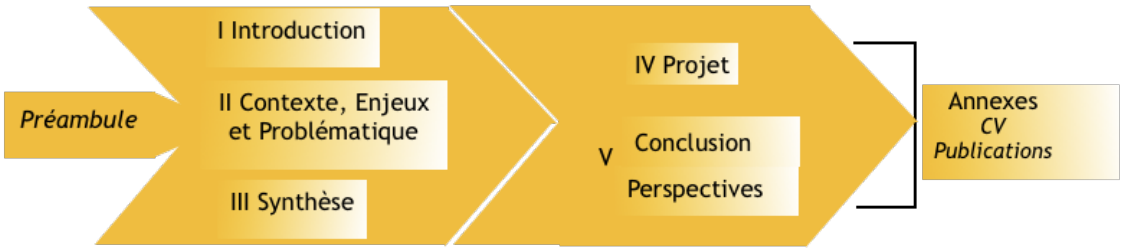


FIGURE 1.1 – Vue d'ensemble

Le mémoire est organisé comme suit :

— **Le chapitre 1- Préambule**. Celui-ci se poursuit en décrivant brièvement mon parcours universitaire et professionnel afin que le lecteur ait une vision globale de mon travail. J'y décris également une synthèse de mes responsabilités d'encadrement, des collaborations et des financements obtenus dans la première partie de ma carrière.

— **Le chapitre 2- Introduction** explicite les divers paramètres qui ont permis de définir et cadrer le projet qui a émergé de l'ensemble des activités et réflexions issues de cette diversité.

— **Le chapitre 3- Contexte scientifique, Enjeux et Problématique** rappelle tout d'abord les concepts de base de la biologie moléculaire et de la génétique, nécessaires pour faciliter la compréhension des travaux de recherches présentés par la suite. Il rappelle ensuite les défis informatiques liés à l'intégration de données et l'extraction de connaissances biologiques que j'ai tenté de relever et qui constituent les piliers de mon travail de recherche.

— **Le chapitre 4- Synthèse** présente la synthèse de mes travaux de recherche qui ont débuté durant ma thèse et se sont poursuivis au cours de mes différentes expériences professionnelles. Ainsi, je présente les propositions et réalisations relatives aux architectures de médiation et

orientées services. Seront détaillés les travaux d'encadrement doctoral sur la ré-écriture de requêtes SPARQL-SQL et la création automatique de services Web. Je présente ensuite les travaux effectués pour le passage à l'échelle dans la recherche d'information génomique et ceux relatifs à ma contribution au portail d'ontologies pour l'agronomie : AgroPortal. Enfin, je présente mes travaux récents sur l'intégration de données et la gestion de connaissances dans le domaine agronomique. Pour cela, je décris la base de connaissance AgroLD qui utilise les technologies du Web sémantique pour représenter et gérer l'information biologique.

— **Le chapitre 5- Projet** détaille les défis auxquels les domaines de la biologie moléculaire des plantes et de l'agronomie font face. Dans ce chapitre, je propose ma vision pour le développement de méthodes permettant de capitaliser sur les données acquises et en extraire les connaissances afin de répondre aux questions soulevées par les chercheurs du domaine.

— **Le chapitre 6- Conclusion et perspectives** conclut ce mémoire en lien avec mon expérience passée et apporte des éléments de discussion sur le projet de recherche que je propose de mener dans un environnement inter-disciplinaire biologie et informatique.

— **Le chapitre 7- Curriculum Vitae** est la première annexe du mémoire. Il présente mon CV en détaillant mon parcours universitaire et professionnel, mes activités de recherche en termes d'animation, encadrements et projets, mes activités d'enseignements. Enfin, il présente une liste exhaustive et une analyse de mes publications.

— **Le chapitre 8- Publications** est la deuxième annexe du mémoire et correspond à la liste des publications sélectionnées dans les chapitres 3, 4 et 5.

1.2 Déroulement de carrière

J'ai un parcours scientifique atypique qui s'est construit sur un cursus universitaire alternant formations diplômantes et expériences professionnelles. Mes premiers contacts avec le monde de la recherche scientifique date de 1998. Après une maîtrise de biochimie, j'ai eu l'opportunité de travailler sur des protocoles expérimentaux en biologie moléculaire au sein d'équipes de recherche de l'INRA. En 1999, j'ai poursuivi ces expériences professionnelles au CIRAD (UMR PIA), pour développer et analyser une banque de marqueurs micro-satellites chez le cacao. J'ai pu alors constater l'importance de l'informatique pour la gestion et le traitement des données à l'échelle de la biomolécule. Un tel constat m'a conduit à compléter ma formation de biologiste avec une année de DESS en informatique. J'ai pu alors aborder les problématiques associées à la structuration et au traitement des données moléculaires sous un angle nouveau lors du stage de fin de cursus du DESS en 2000, qui s'est déroulé dans la même unité de recherche.

Mon parcours scientifique (cf. schéma 1.2) a démarré en 2001 suite à l'obtention du cursus universitaire DESS informatique, en tant qu'ingénieur d'étude en bioinformatique (IE2) dans le groupe d'E. Guiderdoni au CIRAD (UMR PIA) et dans le contexte du projet ANR Génoplante « Analyse fonctionnelle du génome du riz : création d'une collection de 15.000 lignées de mutants d'insertion de riz ». L'objectif de ce projet était de créer et caractériser une collection de mutants génétiques chez la variété *Oryza sativa sous-espèce nipponbarre*. Le projet était très ambitieux sur le plan informatique et comprenait tous les aspects de traitements de séquences génomiques mais aussi l'informatisation des processus d'analyse phénotypique (ce que l'on appelle aujourd'hui le phénome). Un des défis du projet portait sur la mise en place d'un système intégré, pouvant répondre aux attentes de différentes équipes de recherche localisées sur divers sites géographiques. Très rapidement, mes activités ont abordé des problèmes, qui au-delà de la haute technicité, impliquaient des réflexions méthodologiques liées à la gestion de données et de connaissances hétérogènes. C'est dans ce contexte que j'ai effectué ma thèse. J'ai bénéficié de l'encadrement de Thérèse Libourel (Pr. LIRMM, Univ. Montpellier, directrice), d'Isabelle Mougenot (Mdc LIRMM - Univ.

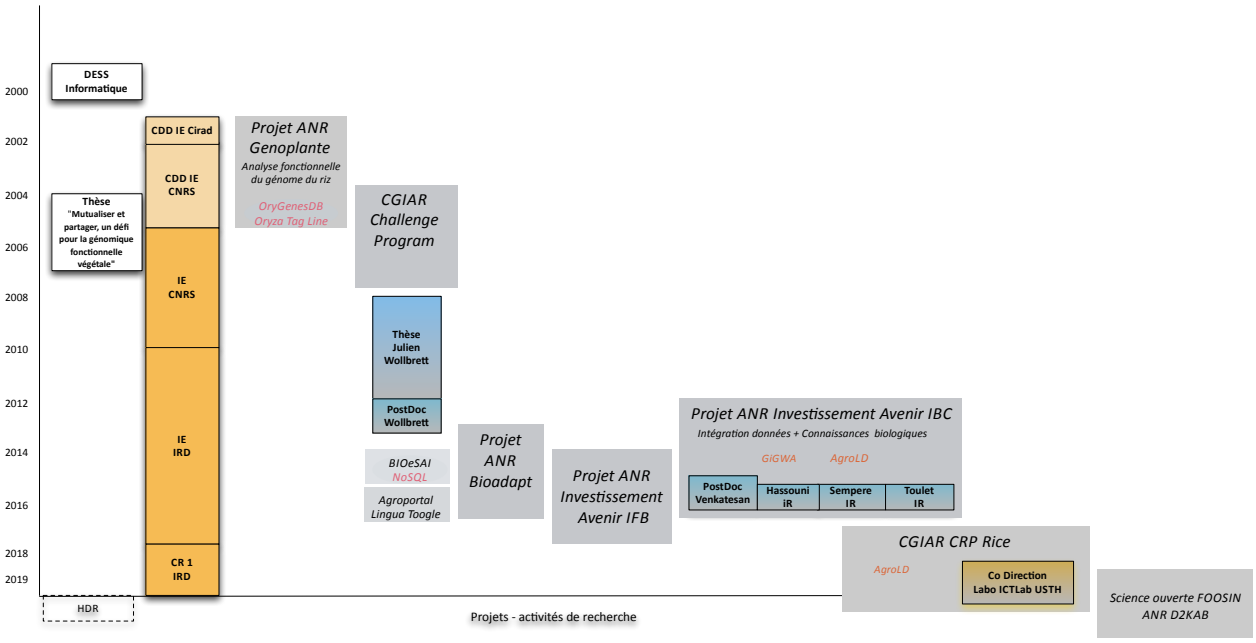


FIGURE 1.2 – Schéma général du parcours

Montpellier) et Manuel Ruiz (Chercheur Cirad). « Mutualiser, partager » des données et connaissances hétérogènes et réparties constituait un sujet important. La proposition effectuée consistait à définir une structure de médiation (paradigme médiateur/adaptateur) reposant sur schéma global permettant une consultation unifiée des différentes sources de données hétérogènes dans le contexte de la génomique fonctionnelle. Le médiateur mis en place s'appuyait sur l'approche GAV (Global As View) fonctionnant comme un schéma global sur un ensemble de vues conçues sur les schémas des sources de données. J'ai obtenu un poste d'ingénieur d'études CNRS dans l'équipe Intégration des Données (ID) dirigée par M. Ruiz au CIRAD (UMR AGAP) quelques temps avant de soutenir ma thèse. Puis, j'y ai poursuivi mes recherches dans cette voie autour des méthodes d'intégration de données dans le domaine agronomique. Je me suis d'abord orienté sur le développement de méthodes automatisant la création d'adaptateurs sémantiques et la formulation de requêtes pour des bases de données biologiques en co-encadrant la thèse de Julien Wollbrett (voir section 4.2). Je me suis, par ailleurs, impliqué dans l'animation de la plate-forme bioinformatique SouthGreen.

Par la suite, j'ai quitté le CNRS en 2010, pour effectuer une mobilité au sein de l'IRD dans l'équipe Génome et Développement du Riz (UMR DIADE) dont les enjeux en matière de partage et d'intégration de données génomiques étaient particulièrement motivants. Je me suis également fortement impliqué, avec d'autres ingénieurs, dans la structuration de la plate-forme bioinformatique naissante transversale «i-Trop», dédiée à plusieurs plusieurs unités IRD.
Afin de répondre aux besoins de gestion des masses de données produites par le séquençage et le phénotypage de nouvelles variétés de riz, j'ai développé des méthodes d'intégration et de stockage basées sur des architectures distribuées détaillées en section 4.3.
Entre 2012 et 2017, j'ai été impliqué dans le projet «Institut de Biologie Computationnelle» (IBC). IBC est un projet ANR «investissement d'avenir» en bioinformatique dont l'objectif était de développer de nouvelles méthodes et logiciels pour le traitement des grandes masses de données biologiques avec des applications dans les domaines de la santé, l'agronomie et l'environnement. Jusqu'en 2017, j'ai été co-responsable de l'axe «intégration des données et connaissances biologiques» qui reprenait les problématiques d'intégration de données pour la biologie des plantes.

Je me suis fortement impliqué dans cette tâche car les problématiques sont très importantes pour l'unité DIADE. Pour mener à bien cette coordination, j'ai partagé mon temps entre les locaux d'IBC situé au LIRMM et l'équipe RICE. J'ai pu ainsi créer une synergie entre experts de différents domaines de l'informatique et de la biologie afin d'avancer sur des points tels que la gestion des données phénomiques et la gestion des données NGS (Next Generation Sequencing). Mes activités de recherches menées dans le cadre du projet IBC sont décrites en section 4.4.
Depuis le mois de septembre 2016, je travaille, mandaté par l'IRD, en expatriation au Vietnam pour développer *in situ* des approches bioinformatiques avec les partenaires Vietnamiens du LMI RICE [1], afin de permettre une meilleure exploitation de leurs données. J'ai également partagé mon temps en travaillant dans le laboratoire informatique IRD-USTH (Université des Sciences et Techniques d'Hanoi). Fin 2017, j'ai pris la responsabilité du laboratoire informatique en co-direction avec un chercheur Vietnamien. Le laboratoire compte neuf jeunes enseignants chercheurs Vietnamiens avec qui je collabore sur certains aspects méthodologiques. Par ailleurs, je consacre une partie de mon temps à l'enseignement en Master (Informatique et Biologie) ainsi qu'à l'encadrement d'étudiants. Je développe également des collaborations avec l'International Rice Research Institute (IRRI) qui est un des centres du Consultative Group on International Agricultural Research (CGIAR) sur le Riz basé aux Philippines.
Depuis mon entrée à l'IRD en 2010, j'ai eu 3 opportunités de présenter des concours internes d'ingénieur de recherche (IR). Sans succès, car le nombre de postes était très faible et souvent regroupait l'ensemble des domaines scientifique et administratif. Cependant, en 2016, une commission scientifique dédiée aux « Sciences des données et du numérique, CSS5 » a été créée afin de mettre en œuvre une politique scientifique dans ce domaine, aidée tout les ans, par la création de postes de chercheurs. J'ai obtenu un poste de chargé de recherche en 2018, en présentant dans cette commission les thématiques que je décris dans ce mémoire.

1.3 Soutiens financiers et personnes impliquées

Depuis la fin de ma thèse, j'ai pu obtenir des financements pour développer ma recherche. Il s'agissait souvent de financements liés à des tâches précises dans des projets plus importants. Le montant global obtenu par ces projets s'élève à peu près à 400k€. Le tableau 1.1 présente une sélection de projets financés.

Projets	Programme	Date	Montant	Type de soutien	Collaboration	Sujet
GCP Pantheon (M. Ruiz)	CGIAR	2004-2008	(15K€) 400K€	Stages, fonctionnement	IRD, CIRAD, CGIAR	BioSemantic,Intégration de données et métadonnées
IFB Plant Node (J-F Gibrat)	ANR PIA 2011	2012-2017	(120K€) 400K€	Ingénieur (36 mois), fonctionnement	INRA, IRD, CIRAD, CNRS	AgroLD, intégration de données
IBC (O.Gascuel)	ANR PIA 2011	2012-2017	(120K€) 2.842M€	PostDoc (24 mois), stages, fonctionnement	IRD,INRA, CIRAD, CNRS, UM	AgroLD, Gigwa et intégration de données
PostDoc Numev (P. Larmande)	LABEX NUMEV	2015-2016	(50K€)	PostDoc (18 mois)	IRD, CNRS, UM	Bioinformatique, pipeline, NGS
BIOeSAI (P. Larmande)	IRD Spirale	2014-2015	(11K€)	Stages	IRD	Bioinformatique, intégration de données
CRP-RICE (A. Ghesquiere)	CGIAR	2017-2022	(60K€) 1,5 M€	Stages	IRD, CIRAD, centres CGIAR	Bioinformatique,AgroLD, intégration de données
D2KAB (C. Jonquet)	ANR generic call 2018	2019- 2023	(12K€) 950K€	Stages fonctionnement	INRIA, INRA, IRSTEA, CNRS-CEFE,IRD	AgroLD, Linked Data
FOOSIN (S. Aubin)	ANR generic call 2018	2019- 2023	(16K€) 76K€	Stages, fonctionnement	INRIA, INRAE CNRS,IRD	AgroLD, FAIR Data

TABLE 1.1 – Sélection de projets financés. La majorité des budgets obtenus a été allouée à l'IRD. Le premier budget entre parenthèse correspond à la somme perçue alors que le budget global est indiqué à la suite.

1. https://sites.google.com/site/lmiricevn

Les travaux présentés dans ce mémoire ont été réalisés soit directement par moi, soit par une personne que j'ai pu encadrer ou co-encadrer avec un autre collègue permanent. Le tableau 1.2 contient une sélection d'encadrements effectués au cours de mon expérience professionnelle.

Personne	Status	Date	Projet	Type de soutien	Co-supervision	Situation suivante
Sébastien Fromentin	M2	2007	Thèse	CIRAD-AGAP		SS2I bioinformatique
Julien Wollbrett	M2	2008		CIRAD-AGAP	M.Ruiz	Doctorant
Julien Wollbrett	Doctorant	2008-2011	BioSemantic	CIRAD-AGAP	M.Ruiz	PostDoc CIRAD
Julien Wollbrett	Post-Doctorant	2012-2013	BioSemantic	CIRAD-AGAP	M.Ruiz	PostDoc CNRS (Roscoff)
Florian Philippe	M2	2014	Gigwa	IBC	G.Sempere	Ingenieur INRA (URGI)
Aravind Venkatesan	M2	2014–2016	AgroLD	IBC		Bioinformaticien EBI (Uk)
Anne Toulet	Ingénieur	2015-2017	AgroPortal	IBC	C. Jonquet	Post-Doctorante Labex Numev
Nordine El Hassouni	Ingénieur	2015-2017	AgroLD	IFB	M. Ruiz	Start-up Data Science
Imène Chentli	M2	2015	AgroLD	IBC	K. Todorov	Ingénieur CNRS
Gildas Tagny	M2	2015	AgroLD	IBC	A. Venkatesan	Doctorant Mines Ales
Sara Remini	M2	2016	AgroLD	IBC	K. Todorov	CDD Data Science
Ahmed Sayadi	M2	2018	AgroLD	IRD	K. Todorov	
Serge Sonfac	M2	2019	AgroLD	IRD	K. Todorov	Doctorant Univ. Tarbes

TABLE 1.2 – Sélection d'encadrements réalisés.

1.4 Collaborations

De par la nature pluri-disciplinaire de mes travaux de recherche, j'ai toujours développé des collaborations aussi bien avec des biologistes qu'avec des informaticiens. J'ai eu l'occasion de développer de nombreuses collaborations tant sur le plan local, avec des équipes de mon unité et d'autres unités Montpelliéraines, que sur le plan national et international par le biais de programmes transversaux ou de projets internationaux.

A l'IRD : j'ai eu l'occasion de collaborer avec de nombreux collègues sur l'exploitation de leur données et l'intégration avec des données externes. Au sein de l'unité DIADE, j'ai collaboré avec l'équipe RICE (L. Albar, A. Ghesquière, M. Lorieux, F. Sabot et C. Tranchant) sur l'exploitation de données expérimentales dans le cadre des projets CRP-RICE et LanPanTOGGLE. J'ai également collaboré avec M. Mirouze sur la visualisation de données épigénétiques. Avec l'équipe EDI, j'ai pu formaliser un modèle de représentation de connaissances dans le cadre du projet BioESAI. Dans le projet ANR Africrop, en collaboration avec l'équipe Dynadiv (P. Cubry, Y. Vigouroux), j'ai pu mettre en œuvre l'application Gigwa pour gérer d'importantes quantités de données génomiques. J'ai également collaboré avec d'autres unités IRD, notamment avec IPME dans le cadre de formation en bioinformatique (A. Dereeper et S. Cunnac) et au niveau de projets de recherche transversaux (CRP RICE).

Au niveau de Montpellier : j'ai des collaborations de longue date avec des membres de l'unité AGAP. Particulièrement avec l'équipe bioinformatique « Intégration de données » (M. Ruiz, G. Droc, G. Sempere), pour le développement de méthodes et d'applications, mais également avec l'équipe DAR (E. Guiderdoni, C. Perin, A. Dievart) sur l'exploitation de données de riz. Dans le cadre du projet IBC, j'ai eu des collaborations avec des membres de l'unité INRA MISTEA (P.Neveu), LIRMM FADO (K. Todorov), LIRMM SMILE (C. Jonquet) et INRIA Zénith (P. Valduriez) sur le développement de méthodes d'intégration de données massives et d'extraction de connaissances.

Au niveau national : Travaillant sur les thématiques en intégration de données biologiques, j'entretiens, depuis longtemps, des collaborations avec l'unité INRA URGI (F. Legeai et D. Steinbach (2004-2010), C. Pommier, M. Alaux, H. Quesneville). Nous avons collaboré sur des projets ANR Génoplante, IFB, D2KAB et FOOSIN. Nous sommes également partenaires dans le cadre du réseau européen de plate-formes bioinformatiques Elixir. J'ai eu l'occasion de collaborer avec l'équipe INRIA WIMMICS dans la mise en place d'un prototype d'intégration de données utilisant leurs outils. Enfin, j'ai collaboré avec des membres du LRI équipe

bioinformatique (S. Cohen-Boulakia et C. Froidevaux) sur les workflows scientifiques dans le cadre du projet IBC.

Au niveau international : Dans le domaine de la bioinformatique et de l'intégration de données, j'entretiens une collaboration depuis 2008 avec l'équipe bioinformatique de l'IRRI dans le cadre des projets GCP Pantheon et CRP Rice. Plus récemment, à travers le projet IFB, j'ai été connecté au réseau Européen de bioinformatique Elixir et eu l'occasion de collaborer avec les développeurs du logiciel d'alignement d'ontologies AML (D. Faria). j'ai été également impliqué comme expert dans le groupe de travail RDA Wheat data Interoperability afin de mettre en œuvre un cas d'usage pour AgroLD[2]. Actuellement, dans le cadre de mon affectation au Vietnam, je collabore avec les chercheurs du LMI RICE et d'ICTLab. J'ai également développé des collaborations autour de la représentation de connaissances et de l'intégration, avec des instituts de recherche Japonais : le DBCLS (S.Goto, J-D Kim), le NARO (H. Kanegae, T. Tanaka), le NIG (Y. Sato).

2. https://www.rd-alliance.org/groups/wheat-data-interoperability-wg.html

Chapitre 2

Introduction

L'agriculture jouera un rôle déterminant dans les années avenir face à l'augmentation importante de la population mondiale et au changement climatique que nous vivons. Ce dernier aura des impacts importants dans le domaine agricole, à la fois un impact direct sur la productivité et un impact indirect tel que le changement dans la disponibilité de l'eau, les parasites et les maladies, et l'utilisation des terres. La dernière évaluation du Groupe d'experts intergouvernemental sur l'évolution du climat (GIEC) a confirmé que ces impacts risquent d'être gravement ressentis dans les pays en développement (GIEC, 2018) et qu'ils auront un fort effet sur les petits producteurs et les personnes à faible revenus [12].

Le riz est la première céréale mondiale en terme de production pour l'alimentation humaine. Cultivée dans les zones tropicales qui sont elles même soumises à de fortes contraintes liées au changement climatique, son agriculture posera un enjeu majeur dans les années futures. Pour relever les défis de la croissance alimentaire mondiale dans ce contexte, il est crucial d'améliorer les capacités de production notamment par l'amélioration génétique des plantes.

La biologie est une science expérimentale visant à élucider le fonctionnement du vivant et chercher à déterminer quelles sont les lois qui le régissent. Aujourd'hui, il est établi que le génome en est l'élément central autour duquel gravite de nombreux paramètres et facteurs influençant son expression. Le phénotype correspond à l'expression « visible » du génome (et des gènes qui en sont les éléments unitaires) et par définition qui est quelque chose que l'on voit (que l'on peut observer de manière tangible). Pour un biologiste, valider une hypothèse de recherche nécessite la mise en place de nombreuses expérimentations et la production d'autant d'observations, de résultats ou de données (selon l'angle dans lequel on se place).

Récemment, les progrès des technologies de séquençage et des méthodes de phénotypage à haut débit ont révolutionné l'analyse du vivant. Elles sont utilisées par les scientifiques pour déchiffrer la complexité des systèmes biologiques. Les résultats souvent matérialisés par plusieurs centaines de gènes qu'il faut analyser pour en identifier seulement qu'une fraction associée à une maladie ou à un phénotype étudié. À un certain stade, chaque scientifique doit choisir les gènes à étudier de manière plus approfondie en laboratoire. Souvent, ce choix est subjectif, car il est basé sur des connaissances partielles des interactions entre le gène et le phénotype.

Premier point de jonction avec le domaine informatique, les bases de données constituent une source majeure de connaissances pour la recherche en sciences de la vie. Actuellement, il existe plus de 2 000 systèmes de bases de données et d'information disponibles via Internet, qui représentent ces données moléculaires [176]. Chaque année, de nouvelles bases de données et systèmes d'information utilisables via Internet apparaissent. Une caractéristique de tous ces systèmes, est que leurs mises à jour tant sur les données que sur le système sont constantes. Une autre difficulté majeure est que ces sources de données sont trop nombreuses pour être toute identifiées et

utilisées par les biologistes. A ce stade, il existe peu d'outils permettant de connaître leur existence et de recommander leur utilisation en fonction de leur contenu. Depuis quelques années des projets de référencement utilisant des métadonnées ont été développés comme par exemple BioMOBY [229], BioCatalogue [19] et plus récemment FAIRSharing [183]. Toutefois, de nombreuses connaissances résident également sous d'autres supports tels que les publications scientifiques, les site Web ou les réseaux sociaux (ex : Twitter, Facebook, etc.).

Une meilleure compréhension des relations gène-phénotype nécessite une intégration de données biologiques de diverse nature. Or, les différents points abordés précédemment contribuent à la principale raison pour laquelle le processus automatique d'accès aux données reste difficile même si les données sont disponibles sur Internet. Pour les biologistes, l'inspection manuelle de ces ressources disponibles sur Internet est une tâche fastidieuse pour laquelle des méthodes informatiques doivent être appliquées. En effet, il n'est pas facile d'interroger ces données et d'avoir une réponse claire tant la masse d'information est difficile à gérer. Aujourd'hui encore le développement d'outils permettant un accès à ces données distribuées et hétérogènes constitue une partie importante de la recherche en bioinformatique et ce depuis plus de 20 ans. Cette thématique a permis le développement d'une importante communauté dont DILS [1], SWAT4LS [2] et BioHackathon [3] sont les plus emblématiques conférences et des infrastructures de recherche comme Elixir [4] ou des instituts de recherche comme IFB [5] sont des éléments structurants.

Ce manuscrit n'a pas vocation d'aborder tout ce qui constitue la richesse des recherches évoquées, mais plutôt de donner un regard personnel sur comment, au travers d'une expérience passée depuis 20 ans dans ces domaines, une direction de recherche a émergée comme un fil rouge que j'aimerai poursuivre et que j'illustrerai à travers différents projets déjà initiés ou en perspectives.

Mon projet de recherche aborde le problème de **la représentation de la complexité des données biologiques afin d'en extraire de la connaissance permettant d'identifier les mécanismes moléculaires contrôlant l'expression de phénotypes** chez les plantes.

L'objectif de ce projet sera de déterminer si la représentation d'information sous forme de graphes de connaissances est adaptée pour formuler des hypothèses de recherche permettant de lier le génotype au phénotype. En prenant le riz comme modèle, il s'agira de construire des réseaux d'interactions moléculaires entre gènes à partir de données éparses (articles scientifiques, bases de données publiques, données expérimentales, etc.) afin d'identifier les gènes clés pour l'amélioration des plantes.

Dans un premier temps, afin d'aborder la question initiale, comment intégrer ces données diverses pour faciliter l'identification de gènes importants pour les biologistes et leur analyse, plusieurs approches seront envisagées : **intégration dynamique des données, annotation sémantique, extraction des connaissances, enrichissement des connaissances, priorisation de gènes candidats.** Elles sont résumées dans la figure 2.1.

1. Une première voie consistera à transformer ces données selon les principes FAIR [230] et les **intégrer dynamiquement** dans une base de connaissances pour les rendre plus facilement réutilisables et exploitables automatiquement. Une attention particulière sera portée aux

1. Data Integration for Life Science - `https://link.springer.com/conference/dils`
2. Semantic Web for Life Science - `http://www.swat4ls.org`
3. `http://www.biohackathon.org`
4. `https://elixir-europe.org`
5. Institut Francais de Bioinformatique - `https://www.france-bioinformatique.fr`

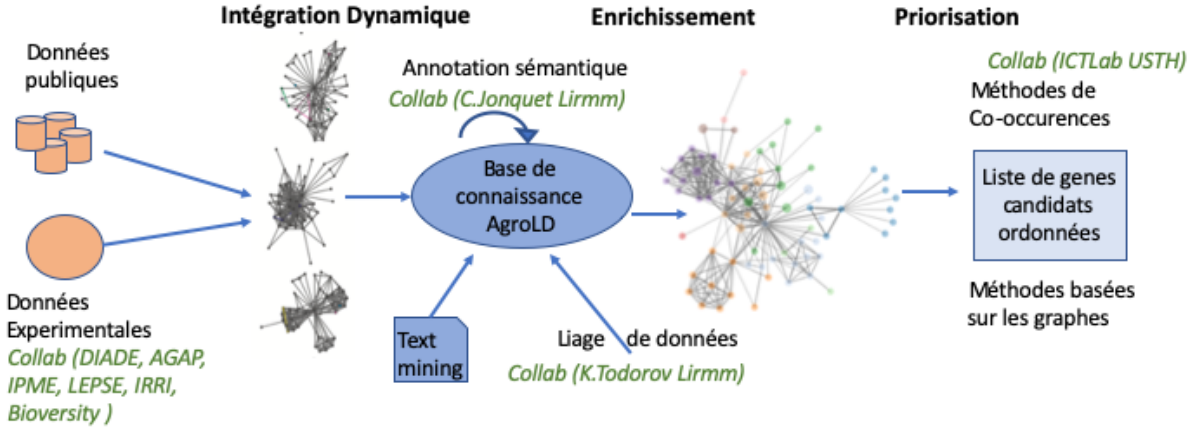


FIGURE 2.1 – Schéma général du projet de recherche

données expérimentales produites par les chercheurs des unités plantes Montpelliéraines et les partenaires internationaux. S'agissant souvent de données massives (i.e., données de génotypages ou d'images de phénotypes), la méthode proposée évitera de transformer intégralement les données afin de garantir de bonnes performances.

2. Une deuxième voie consistera à proposer de nouvelles méthodes d'**enrichissement** des connaissances.
   - Dans un premier temps, se focaliser sur des **méthodes d'annotation sémantique** pour lier plus facilement différentes sources de données avec les concepts ontologiques et en extraire des informations. Nous utiliserons de nouvelles fonctionnalités d'AgroPortal, portail d'ontologies de référence pour le domaine agronomique.
   - Dans un deuxième temps, comme une grande partie de la connaissance réside dans les données non structurées telles que les publications scientifiques, les champs texte des bases de données ou des réseaux sociaux, le développement de **méthodes de fouille de texte** pour identifier les entités biologiques et leurs relations sera envisagé.
   - Toujours dans l'objectif d'améliorer les analyses portant sur les données biologiques, l'idée d'utiliser les informations extraites par les différentes approches précédentes pour « lier » les données sous forme de graphes de connaissances est la dernière voie d'enrichissement envisagée. Ainsi, pour enrichir les liens entre les différents graphes générés et produire un réseau d'interactions qui permettra la découverte de nouvelles connaissances, **de nouvelles méthodes de liage de données RDF** spécifiques aux problématiques bioinformatiques seront développées.
3. Enfin, la recherche d'information parmi ces graphes nécessite le développement de méthodes pour trier pertinemment les résultats. La priorisation de gènes candidats permet d'identifier et de classer parmi un grand nombre de gènes, ceux qui sont fortement associés au phénotype ou la maladie étudiée. Plusieurs **méthodes et algorithmes de priorisation de gènes candidats** seront évalués et proposés.

Chapitre 3

Contexte scientifique, Enjeux et Problématique

Dans ce chapitre nous allons rappeler tout d'abord, les concepts de base de la biologie moléculaire et de la génétique, en les approfondissant dans le cas du riz pour faciliter la compréhension des travaux de recherches présentés par la suite. Puis nous aborderons les enjeux computationnels liés aux divers progrès de la biologie moléculaire avant de détailler les problématiques que nous avons abordées au confluent de la biologie et de l'informatique ainsi qu'un état de l'art autour des principales avancées existantes.

3.1 Biologie moléculaire et génétique

La biologie est une science expérimentale visant à élucider le fonctionnement du vivant et chercher à déterminer quelles sont les lois qui le régissent. Aujourd'hui, il est établi que le génome en est l'élément central autour duquel gravite de nombreux paramètres et facteurs influençant son expression.

Le dogme central de la biologie moléculaire (Figure 3.1) suggère que tous les processus biologiques d'un organisme proviennent des informations codées dans son ADN génomique. De fait, décrypter la séquence complète du génome (i.e. la totalité de l'ADN répartie sur les chromosomes) permettrait de comprendre l'ensemble des mécanismes biologiques. Nous ne rentrerons pas dans le détail des aspects de la biologie moléculaire. Toutefois, nous renvoyons le lecteur vers des livres de vulgarisation tels que « The Molecular Biology of the Cell » [1]

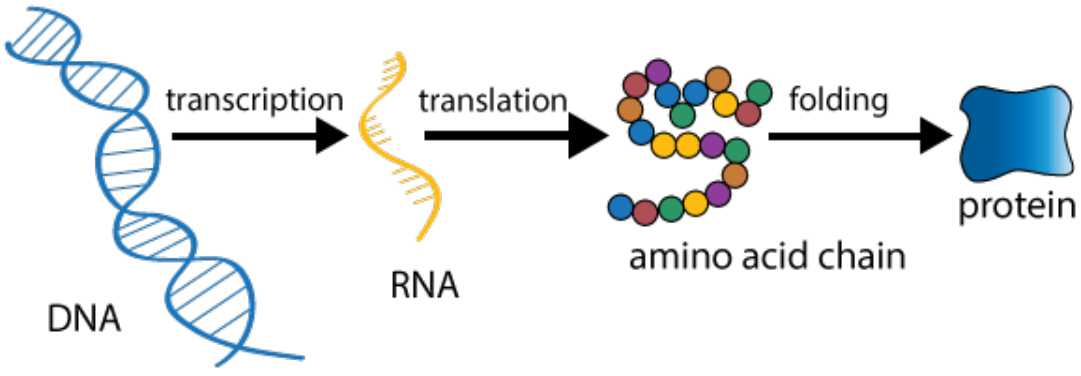


FIGURE 3.1 – Le dogme central de la biologie moléculaire indique que l'information contenue dans l'ADN génomique est successivement transformée en ARN, chaîne amino acides et protéine. Crédits :biosocialmethods.isr.umich.edu

1. https://epdf.pub/molecular-biology-of-the-cell-5th-edition.html

Le phénotype correspond à l'expression «visible» du génome (et des gènes qui en sont les éléments unitaires) et par définition qui est quelque chose que l'on voit. Pour un biologiste, valider une hypothèse de recherche nécessite la mise en place de nombreuses expérimentations et la production d'autant d'observations, de résultats ou de données (selon l'angle dans lequel on se place). Récemment, les progrès des technologies de séquençage et des méthodes de phénotypage à haut débit conduisent à une explosion de données. Elles sont utilisées par les scientifiques pour déchiffrer la complexité des systèmes biologiques et comprendre comment les phénotypes sont structurés au niveau moléculaire. Les résultats sont souvent matérialisés par plusieurs centaines de gènes qu'il faut analyser pour identifier seulement une fraction de gènes associés au phénotype étudié. À un certain stade, chaque scientifique doit choisir les gènes à étudier de manière plus approfondie en laboratoire. Souvent, ce choix est subjectif, car il est basé sur des connaissances partielles des interactions entre le gène et le phénotype.

3.1.1 Caractérisation des relations génotype-phénotype

La compréhension des relations génotype-phénotype est un des axes les plus importants de la recherche, tant en santé humaine avec des applications sur la prédiction des risques ou le traitement thérapeutique, que pour les animaux et les plantes pour accélérer la reproduction des caractères importants pour la production agricole. Or, les interactions génotype-phénotype sont complexes à identifier. Au cours des dernières années, une multitude d'études GWAS (*Genome Wide Association Studies*) ont identifié de nombreux variants génétiques associés à des maladies complexes ou à des caractères phénotypiques. Toutefois, même si ces découvertes enrichissent grandement nos connaissances sur les bases génétiques de la variabilité phénotypique, la plupart des variations identifiées jusqu'à présent n'expliquent qu'une faible proportion des facteurs génétiques causaux, laissant à découvrir et expliquer l'héritabilité restante [136]. Par ailleurs, même avec une compréhension complète de la génétique d'un caractère phénotypique complexe, la prédiction des variations phénotypiques (e.g. expliquer un changement de couleur ou de taille du grain) reste encore difficile à expliquer. Une des raisons est que la majorité de ces variations génétiques liées à une maladie ou à un trait (ou caractère phénotypique) se trouvent dans des régions non codantes du génome, ce qui complique leur annotation fonctionnelle et représente un des plus grands défis de l'ère «post-GWA» [67, 90].

Lier, à l'échelle du génome, les variants génétiques à la diversité phénotypique, est l'un des objectifs majeur de la biologie. Or, notre compréhension d'une telle cartographie génotype–phénotype ne peut être établie sans données phénotypiques détaillées [91]. Hélas, notre capacité à caractériser les phénomes - l'ensemble des phénotypes d'un individu - est largement en retard sur notre capacité à caractériser les génomes, comme l'ont constaté Furbank et Tester [68]. En conséquence, la phénomique (i.e. phénotypage à haut débit et multi-échelle) émerge comme une discipline combinant de nouvelles technologies d'observation du vivant (i.e. caméra, capteurs, etc.) et permettant d'accélérer les progrès dans notre compréhension de la relation entre génotype et phénotype.

Les relations génotype-phénotype sont aussi très liées/sensibles aux facteurs environnementaux (e.g. la cigarette augmente fortement les risques de cancers, la sécheresse favorise une baisse de production). Ces relations sont souvent conceptualisées :

Génotype (G) + Environnement (E) + génotype x environnement (GxE) -> Phénotype (P)

Ainsi, pour étudier ces interactions de manière reproductible, il est nécessaire de travailler dans des conditions environnementales stables et contrôlées.

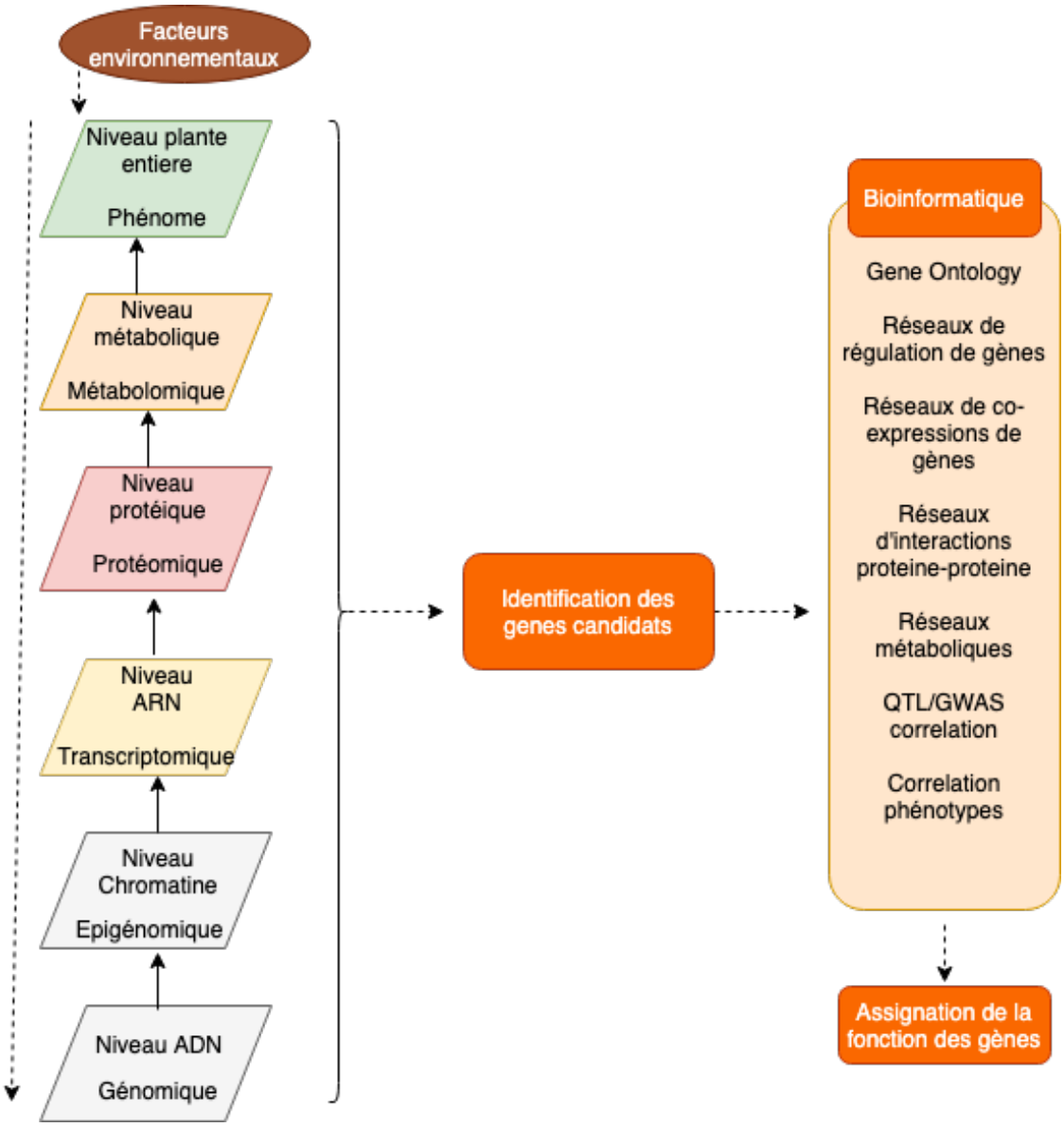


FIGURE 3.2 – Différentes échelles de la régulation de l'expression des gènes conduisant à un phénotype [33]

3.1.2 Les mécanismes qui régulent l'expression des gènes

La génomique ainsi que d'autres technologies d'analyses moléculaires haut débit comme l'épigénomique, la transcriptomique, la protéomique et la métabolomique sont devenues les méthodes d'analyses standards dans ce domaine (que l'on nomme « omique ») et dont l'objectif est d'étudier le système biologique moléculaire entier. Par ailleurs, la phénomique développe des méthodes pour étudier les phénotypes de manière précise et en quantité importante. Comme le montre la figure 3.2, la régulation de l'expression des gènes conduisant à un phénotype peut intervenir à différents niveaux au sein d'une cellule et de l'organisme.

— D'abord, au niveau de **l'ADN génomique**, sur lequel de simples mutations (SNP) ou de grandes modifications de sa structure (délétions, modifications ou insertions de grand fragments appelés CNV) peuvent modifier l'expression des gènes.

— Au niveau de **l'épigénome** - ensemble des propriétés physico-chimiques de l'ADN et des protéines histones sur lesquelles il est enroulé - qui contrôle la structure de la chromatine (complexe ADN-histones structurant un chromosome) et que des facteurs épigénétiques permettent de modifier. L'épigénomique est la discipline qui étudie l'ensemble de ses facteurs et leurs liens avec la structure de la chromatine. L'épigénome est très sensible aux facteurs environnementaux externes qui agissent comme stimuli (positif ou négatif). Comme le montre la figure 3.3 [2], des modifications chimiques de la chromatine permettent de libérer l'accès à l'ADN et favorisent l'expression des gènes.

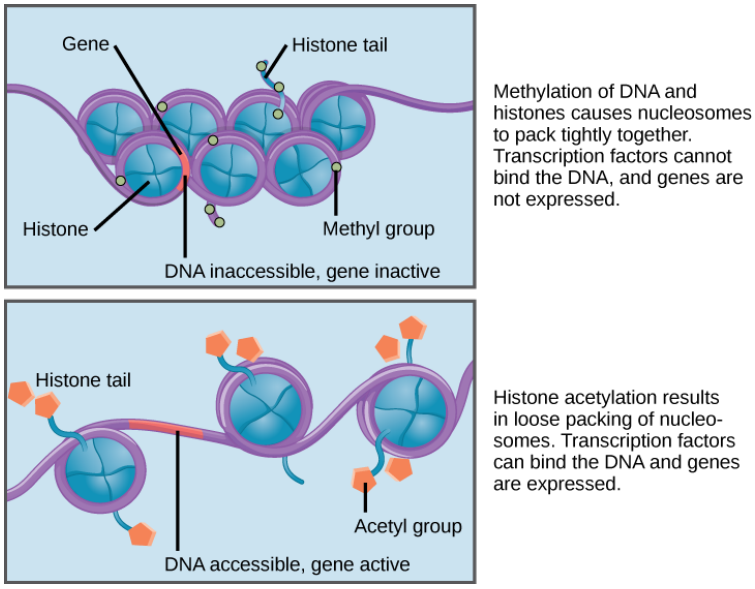


FIGURE 3.3 – Mécanisme d'ouverture du nucléosome - complexe histones-ADN - pour permettre la transcription des gènes. Crédits :Lumen

— **La transcriptomique** fait référence à l'analyse de l'ensemble des molécules d'ARN, de l'ARN codant pour les protéines à l'ARN non codant. Le transcriptome peut s'appliquer à un organisme entier ou à un type de cellule spécifique. Des méthodes actuelles permettant d'identifier de manière exhaustive et ciblée l'expression de presque tout les types d'ARN. L'analyse du transcriptome renseigne directement sur le taux et la dynamique (quantité et variation temporelle) d'expression des gènes, leur co-expression et leur spécificité lié à un type cellulaire ou un tissu. Il permet également des révéler des mécanismes de régulation impliquant les ARN non codant tels que les miRNA, siRNA et leurs familles.

— **La protéomique** est l'étude de l'ensemble des protéines exprimées par un génome, cellule, tissu ou organisme pour un temps donné. Comme pour l'ARN, elles ont un lien direct avec les gènes à partir duquel elles sont traduites. L'action du protéome sur la régulation des gènes est multiple. Les protéines peuvent interagir directement i) sur les gènes pour modifier leur expression (stimuler ou stopper), ii) sur les ARN pour modifier leur expression ou leur stabilité, iii) sur les protéines elles-même en auto-régulation ou interaction, iv) sur le métabolome dont elles sont les acteurs principaux. Les protéines agissent à différents niveaux dans l'organisme et sont impliquées dans tout les processus biologiques.

— **La métabolomique** est l'analyse des petites molécules chimiques présentes dans une cellule, tissu ou organisme. Ces molécules interviennent dans les processus biologiques comme co-facteurs catalytiques. Les réseaux (voies) métaboliques sont des séquences de réactions

2. https://courses.lumenlearning.com

biochimiques impliquant les protéines et ces petites molécules. Ces réseaux peuvent être différents selon l'organisme, les stades de développement, les localisations sub-cellulaires, etc. L'information acquise sur ces réseaux constitue une base importante pour la compréhension de la biologies des systèmes.

— **Le phénome** représente l'ensemble des caractères phénotypiques (traits) observés chez un organisme. Selon certains experts, il peut inclure les dimensions citées précédemment, toutefois en général le phénome fait référence aux observations externes réalisées au niveau de l'individu (e.g. la plante). Outre les traits qui sont principalement déterminés génétiquement (par exemple la couleur des cheveux), de nombreux traits dépendent d'effets environnementaux, tels que les stress biotiques ou abiotiques.

Même si les technologies permettent d'aller toujours plus loin dans l'obtention de nouvelles données, notre connaissance du système reste encore parcellaire pour élucider les mécanismes moléculaires qui régissent l'expression des caractères complexes. Les nouveaux défis consistent à comprendre les relations complexes existant entre le génome, l'épigénome, l'environnement et le phénome. 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 et permettre de prédire les phénotypes.

3.1.3 La biologie moléculaire dans le cas du riz

L' hypothèse du dogme central de la biologie moléculaire, a entraîné dés 1990, le développement de grands projets de séquençage dont le projet international de séquençage du génome du riz (IRGSP) en 1998, regroupant des chercheurs de dix pays, dont des chercheurs Français. Le riz, particulièrement l'espèce Oryza sativa qui est la plus représentative du genre, fut le premier grand projet de séquençage de génome pour les plantes cultivées. Oryza sativa est un génome diploide de type AA qui comprend deux sous-espèces principales (voir figure 3.4) : la variété japonica à grain court et collant, et la variété de riz indica à grain long et non collant. Les variétés Japonica sont généralement cultivées dans le nord-est de l'Asie et dans les zones montagneuses tandis que les variétés Indica sont principalement des riz de plaine, cultivés principalement en immersion, dans les zones tropicales en Asie. Japonica (variété nipponbare) fut le premier génome à être séquencé. Une séquence représentant une couverture de 95% de sa longueur totale de 389 Mega-bases fut achevée en 2004. Cette séquence génomique de haute qualité servit pendant de nombreuses années de modèle aux projets de séquençage d'autres cultures céréalières (le sorgho, le mais, le blé) possédant de grands génomes et des contenus chromosomique complexes [139]. La recherche de gènes *ab initio* (i.e. localiser la position des gènes sur le génome) prédit un total de 37 544 séquences codant pour des protéines et une comparaison avec le génome d'Arabidopsis thaliana révéla que 2 859 gènes de riz n'ont pas été observés auparavant dans cette espèce voisine. La variété Indica fut séquencée presque simultanément mais avec une qualité bien inférieure.

L'annotation du génome (i.e. assigner une fonction aux gènes) est absolument essentielle pour utiliser les informations sur le génome dans les études biologiques. Dans le cas du riz, deux projets concurrents ont produit une annotation différente. La première fut réalisée par le TIGR (The institute of Genome Research) et aujourd'hui gérée par la Michigan State University (MSU) [3]. Alors que les membres de l'IRGSP ont lancé le projet officiel d'annotation du génome (RAP) publiant les données à partir de RAP-DB [4]. Dés lors, les deux systèmes co-existent encore aujourd'hui avec un recouvrement partiel des annotations, ce qui complique la tâche des scientifiques pour l'analyse

3. http://rice.plantbiology.msu.edu/cgi-bin/gbrowse/rice/
4. http://rapdb.dna.affrc.go.jp

de leurs données.

Dans les années qui ont suivi, de nombreuses études de génomique fonctionnelles ont été conduites afin de mieux caractériser la fonction de ces gènes identifiés. Nombreuses de ces études consistaient à désactiver les gènes par ciblages spécifiques (cf.Transgénèse [5]) telles que celles décrites dans la section 4.1 et dans laquelle j'ai participé. A l'issue des ces premières découvertes, les scientifiques ont constaté que l'identification des gènes dans le génome ne suffit pas à expliquer les caractères phénotypiques observés chez la plante. Par ailleurs, les analyses de diversité génétiques réalisées sur des populations de plantes de l'espèce O. sativa, révèlent des différences dans l'expression de gènes et dans la présence-absence de certains d'entre eux. Ainsi, succédant à ces premiers projets de séquençage, le projet OMAP « Oryza Map Alignment Project » fut établi dans le courant des années 2000 avec pour objectif de séquencer et étudier la structure évolutive des génomes diploïdes du groupe AA et BB [232].

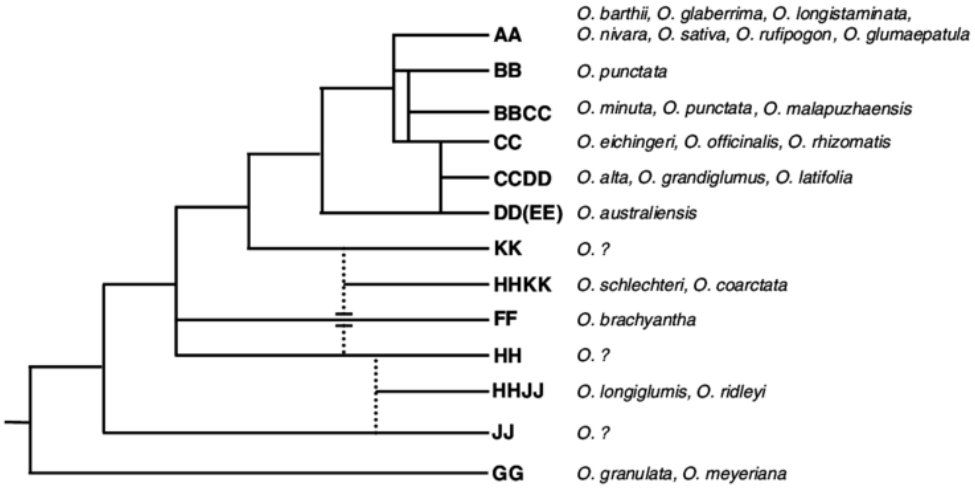


FIGURE 3.4 – Arbre Phylogénétique du genre Oryza (Modifié à partir de Ge et al. 1999). Crédits :Projet OMAP

3.2 Les enjeux computationnels

De nombreux **défis informatique** ont du être relevés afin de permettre l'étude de la diversité à large échelle, chez les êtres vivants, et **l'apport de l'informatique a été important à différents niveaux.** Après avoir rappelé les progrès essentiels réalisés en biologie moléculaire, nous nous focaliserons dans les sections suivantes sur un état de l'art dédié aux grands axes de la recherche informatique qui ont permis les travaux interdisciplinaires précurseurs du projet : **intégration de données, représentation des données, extraction de connaissances**.

3.2.1 La révolution des technologies haut-débit

Récemment, les progrès des technologies de séquençage et des méthodes de phénotypage à haut débit conduisent à une explosion de données. Elles sont utilisées par les scientifiques pour déchiffrer la complexité du système biologique et comprendre les bases moléculaires des phénotypes et des maladies offrant une occasion unique d'accélérer l'amélioration de ce système. Le projet de séquençage à grande échelle le plus récent pour O. sativa est le 3000 Rice Genomes Project [223]. Ce projet, a utilisé une collection de base de 3 000 accessions de ressources génétiques

5. https://fr.wikipedia.org/wiki/Transg%C3%A9n%C3%A8se

de riz, sélectionnées parmi des ressources de l'Institut International de Recherche sur le Riz (IRRI) et de l'Académie Chinoise des Sciences Agricoles (CAAS), et comprenant des accessions provenant de 89 pays répartis en Asie du Sud-Est (33,9%), en Asie du Sud (25,6%) et en Chine (17,6%) incluant des cultivars japonica et indica. Chaque génome des 3 000 accessions contenait des séquences avec une couverture de 14x en moyenne (1x correspondant à une fois le génome), ce qui indique que cette masse de données fournissait une profondeur suffisante pour la détection de polymorphismes mono-nucléotidiques (SNP) fiables. Au total 17 To de données ont été obtenues. D'après une comparaison avec le génome de référence de l'IRGSP-1.0, environ 18,9 M de SNP ont été identifiés. Ces données serviront de ressources fondamentales pour la découverte de nouveaux allèles (i.e. variation d'un gène chez un individu) pour d'importants caractères utiles à l'amélioration du riz et à son adaptation au changement climatique.

Ces recherches visent principalement à comprendre la relation entre génotype et phénotype sur la base d'études d'association pan-génomique (GWAS) et fournissent des informations telles que des polymorphismes génétiques spécifiques pour une variété, la diversité génétique intra et inter population, et des informations sur l'histoire la domestication du riz en Asie.

L'**informatique intervient** d'abord **au niveau du stockage** car d'important volume de données sont générés et doivent être traités. La communauté bioinformatique a également fait de gros efforts sur la formalisation de formats de représentation en mettant au point **des standards d'échange de données**. Les traditionnels **pipelines de traitements de données** ne passant plus à l'échelle, les approches « Big Data » ont été adaptées. Enfin, de nombreux **algorithmes et méthodes** ont été développés pour évaluer la qualité et nettoyer les séquences, effectuer des alignements de séquences sur un génome de référence, réaliser un assemblage de génome *de novo*, visualiser des alignements. Nous renvoyons le lecteur vers des articles de synthèse [215, 55, 166].

3.2.2 Analyses de variabilité

Les études GWAS (*Genome Wide Association Studies*) sont des analyses biologiques étudiant les variations génétiques à l'échelle du génome pour un ensemble d'individus et pour un caractère phénotypique donné (trait). Les marqueurs polymorphes les plus couramment utilisés pour les études GWAS sont les polymorphismes de séquence tels que les SNP et les variants structuraux tels que les indels (i.e. insertion ou délétion de nucléotide chez un individu par rapport au génome de référence) et les CNV (i.e. *Copy Number Variation*, éléments de structures répétées). Les études GWAS sont maintenant préférées aux études de génétique d'association traditionnelles telles que les QTL (*Quantitative Trait Loci*) qui utilisent la cartographie par intervalles pour estimer la position sur la carte génétique et l'effet de chaque QTL. Comme l'illustre la figure 3.5, les locus GWAS regroupent souvent plusieurs centaines de gènes qu'il faut analyser pour identifier seulement une fraction de gènes associés au caractère (trait) étudié. À un certain stade, chaque scientifique doit choisir les gènes à étudier expérimentalement en laboratoire. Souvent, ce choix est subjectif, car il est basé sur des connaissances partielles des interactions entre le génotype et le phénotype.

L'informatique a permis l'étude de la diversité à large échelle, chez les êtres vivants. Par exemple, elle a aidé à la recherche de motifs pour l'analyse de séquences ou la détection de variation génomique [92]. Dans cette dernière problématique, l'apport des méthodes d'apprentissage profond a révolutionné la discipline. Citons notamment le logiciel DeepVariant développé par Google [170]

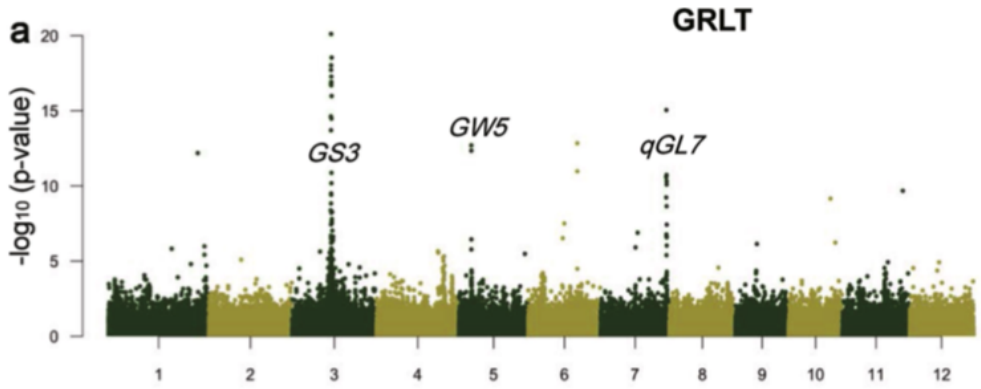


FIGURE 3.5 – Analyse GWAS réalisée pour la longueur du grain (GRLT) chez Oryza sativa (Modifié à partir de Wang et al. 2018). Illustration d'un manathan plot montrant la corrélation entre des variants et le caractère de longueur du grain. Ici chaque point représente un SNP avec sur l'axe des abscisses sa position chromosomique et sur l'axe Y sont degré d'association. Sur cet exemple, les gènes connus ont été indiqués sur les positions et d'autres positions sont potentiellement candidates

3.3 L'intégration de données

Nous avons déjà signalé l'importance de la gestion des données dans le chapitre 2 et les divers écueils rencontrés.

Nous rajoutons de plus, qu'il est important de noter que la qualité des données présentées par ces systèmes de gestion via Internet doit être garantie par chaque fournisseur de données. Les utilisateurs de ces ressources sont très sensibles à la provenance des données et la manière dont elles ont été traitées (ex : informatique, manuelle, etc). C'est un élément important à prendre en compte dans les analyses réalisées à partir de données issues de plusieurs ressources distribuées.

3.3.1 L'hétérogénéité des systèmes

Au-delà de la discussion sur la qualité des données, il est également important de mentionner que ces systèmes sont extrêmement hétérogènes. Dans leur article de synthèse, Leser et Naumann [124] énumèrent une classification des formes d'hétérogénéité existantes que nous avons adapté ici :

— **L'hétérogénéité syntaxique** se retrouve dans le modèle de données (XML, relationnel, objet, graphe, etc.), dans les langages d'interrogation (XQuery, SQL, OQL, SPARQL, etc.), dans les protocoles d'accès (HTTP, etc.), dans les interfaces (REST, SPOAP, .NET, etc.).

— **L'hétérogénéité structurelle** correspond aux différences dans la représentation des données. L'autonomie de conception provoque souvent une hétérogénéité structurelle, schématique et sémantique dans l'intégration des données. L'hétérogénéité structurelle est un cas particulier d'hétérogénéité sémantique, où différents concepts d'un modèle de données décrivent le même problème ou les mêmes données. Ils surviennent quand les schémas de deux sources décrivent différemment un même concept. Par exemple, le nom d'un employé peut être représenté par deux champs *prénom* et *nom* dans une source et par un seul champ *l'identité* dans une autre source. Nous pouvons également citer l'exemple d'un concept défini comme une classe dans une source de données et comme attribut dans une autre.

— **L'hétérogénéité sémantique** caractérise les différences de sens, d'interprétation, de types de termes et de concepts. Les synonymes et les homonymes jouent un rôle majeur dans ces conflits. Les synonymes sont deux mots distincts ayant le même sens. C'est l'exemple de *publication* et *article* qui capturent la même information sur les articles de recherche publiés. Les homonymes sont des mots partageant la même graphie et la même prononciation mais n'ayant pas le même sens. Par exemple, une étoile représentant une planète et l'actrice de cinéma.

D'autres classifications, plus complètes existent, comme celle de Pluempitiwiriyawej et Hammer, «*Classification Scheme for Semantic and Schematic Heterogeneities in XML Data Sources*» [17]

3.3.2 L'évolution des approches d'intégration de données

Les défis majeurs actuels sont liés au développement de méthodes pour l'intégration de ces données hétérogènes et à l'enrichissement de connaissances biologiques.
La figure 3.6 montre que les évolutions des méthodes d'observation du vivant ont bénéficié des avancées technologiques en informatique pour extraire de la connaissance dans les données [219].

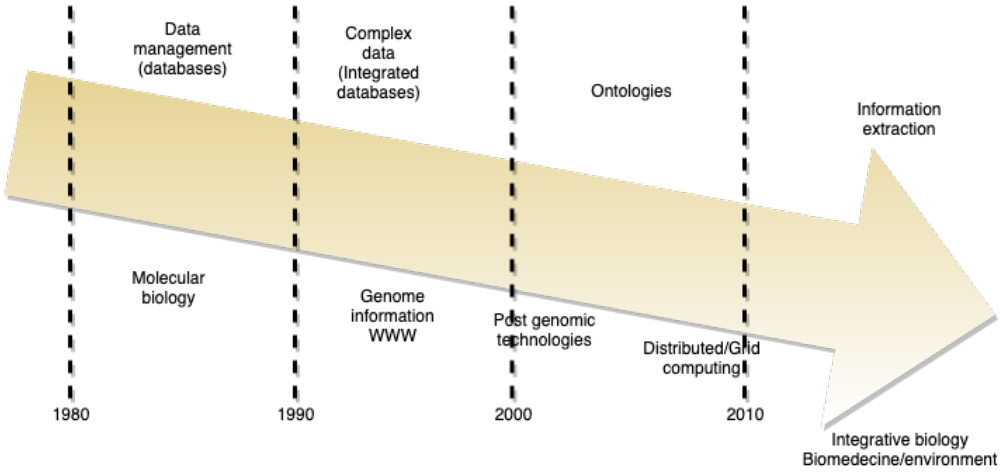


FIGURE 3.6 – Évolution des systèmes d'information en parallèle des méthodes biologiques. Crédits : Valencia 2002 [219]

Le développement de système d'intégration peut devenir extrêmement complexe si le nombre de sources à intégrer est important. En général, les systèmes d'intégration fournissent une vue unifiée de plusieurs sources hétérogènes, autonomes et réparties, facilitant ainsi l'accès à l'information. La méthode est réalisée par l'utilisation d'un schéma global ou d'une ontologie globale, qui fournit une vue réconciliée (consensuelle) des sources locales. Il existe deux approches pour l'intégration : l'intégration matérialisée et l'intégration virtuelle.

L'intégration matérialisée, stocke l'ensemble des données intégrées dans un SGBD en dupliquant ces dernières à partir des sources. Cette approche nécessite de mettre à jour régulièrement les sources et réaliser des extensions du modèle global pour l'ajout de nouvelles sources. Elle présente l'avantage d'avoir des temps d'accès très rapide, car il n'y a pas de communication entre différentes sources de données, ni de limitation des requêtes. En revanche, elle nécessite un stockage volumineux et fiable. Par ailleurs, une étape importante de pré-traitement des données est nécessaire. Le processus d'ETL (Extraction Transform and Load) est la méthode désignée pour intégrer les données dans un SGBD. Elle peut s'avérer très complexe.

L'intégration virtuelle ne stocke pas les données de manière persistante en s'inspirant du modèle de la médiation proposé par Wierderhold [226, 227]. Généralement, les données sont situées sur différents systèmes répartis et interrogées à l'aide d'un schéma global. Un processus d'ETL n'est pas nécessaire, contrairement à l'intégration matérialisée. Les requêtes sont gérées à partir d'un schéma global, tandis que les données sous-jacentes sont «virtuellement» disponibles. La tâche principale du système et d'offrir un protocole d'accès et un langage de requêtes commun à toutes ces sources. Par ailleurs, il doit générer des requêtes complexes pour obtenir, transformer et agréger des données adéquates provenant de différentes sources de données. La communication avec les sources fonctionne généralement à l'aide d'adaptateurs. Leur rôle est d'adapter la requête du médiateur exprimée dans le langage commun au langage de la source, tout en utilisant le bon protocole d'accès. Étant donné que la requête utilisateur est exprimée en fonction du schéma global, une correspondance (ou mapping) entre ce schéma global et les schémas locaux (des sources) est nécessaire afin que les requêtes puissent exécutées par les sources locales. Ce mapping constitue un traitement clé dans le processus général. Il sera utilisé pour réécrire la requête initialement exprimée en fonction du schéma global, en des sous-requêtes exprimées, chacune, en fonction des sources locales.

Deux approches existent pour définir le mapping entre le schéma global et les schémas des sources : *Local As View* (LAV) et *Global As View* (GAV) [81]. Dans l'approche GAV, le schéma global est exprimé à l'aide de vues sur les schémas locaux, à l'inverse de l'approche LAV qui nécessite la description des sources locales en fonction du schéma global. Les approches LAV et GAV ont chacune des avantages et des inconvénients. Ainsi, la LAV favorise l'extensibilité du système d'intégration puisque l'ajout ou la suppression des sources est simple, chaque source étant décrite indépendamment des autres. Mais, la réécriture dans ce cas est un problème complexe. Quant à l'approche GAV, elle favorise, la performance du système quand l'utilisateur pose fréquemment des requêtes complexes puisque les algorithmes de réécriture de requêtes sont plus simples. Cependant, l'ajout ou la suppression d'une source de données nécessite la mise à jour du schéma global pour l'adapter au nouvel état du système. En plus de la LAV et de la GAV, il faut mentionner l'approche GLAV qui est une combinaison des deux approches [122].

Une synthèse des principales approches d'intégration en bioinformatique réalisées au cours des dernières années a été discutée dans Cohen-Boulakia et Leser [39]. Nous en proposons ici une version modifiée et étendue :

— **Les Systèmes de navigation hyper-texte** sont les premières générations de systèmes d'intégration (1985-1995). Ils utilisent comme index, les identifiants d'entités biologiques enregistrées dans des fichiers plats aux formats spécifiques (sans SGBD) ainsi que des liens hyper-texte faisant office de cross-références vers d'autres sources similaires. Surtout un des avantages est qu'ils utilisent des interfaces HTML permettant la recherche et la navigation (SRS [61], Entrez [162]).

— **Les systèmes centralisés gérés par un SGBD et à multi-bases de données** n'ont pas de schéma global. Ces systèmes génèrent de manière interactive des requêtes pour plusieurs bases de données simultanément.

— **Les systèmes de base de données fédérés et les systèmes de type médiateur** sont des systèmes d'intégration virtuels. Ils ne stockent aucune donnée dans un schéma global. Les systèmes fédérés intègrent plusieurs SGBD autonomes dans une base de données fédérée virtuelle unique. En règle générale, chaque base de données est inter-connectée via un réseau informatique ou, dans certains cas, le Web. Par conséquent, les bases de données peuvent

être décentralisées géographiquement (K2/Kleisli [49], DiscoveryLink [79]).

— **Les entrepôts de données** sont des approches d'intégration matérialisées [6]. Ils stockent les données persistantes dans un référentiel de données global, qui est généralement un SGBD relationnel (GUS [49], Atlas [192], BioWarehouse [103], Columba [216]).

— **Les boîtes à outils** qui facilitent la construction de tels entrepôts de données sont très populaires. Parmi elles, citons BioDWH [217] GMOD-CHADO [244], BioMART [197], Intermine [199], Tripal [65, 182, 41].

— **Les systèmes utilisant des ontologies** qui s'appuient sur des schémas à base de graphes pour l'intégration et les requêtes : (TAMBIS (*Transparent Access to Multiple Bioinformatics Information Sources*) [201] et ONDEX [110, 203, 204]).

— **Les Systèmes hybrides** s'appuyant sur plusieurs technologies. Par exemple, Biozon [20] combine une approche SGBD relationnel et une représentation sous forme de graphe.

Toutes ces approches ont le même objectif : fournir des techniques pour surmonter les difficultés liées aux nombreux types des données hétérogènes et fournir un système de recherche aux scientifiques pour appuyer leurs activités de recherche et leurs expériences. Pendant des décennies, les SGBD relationnels ont été majoritairement utilisés pour traiter des données structurées. Cependant, en raison du volume, de la rapidité d'évolution et de la variété des données, ces derniers ne peuvent souvent pas offrir les performances et le temps de latence requis pour gérer des données volumineuses et complexes. L'augmentation de la production de données non structurées issues de capteurs ou technologies haut-débit, pas seulement en biologie, fait émerger de nouveaux besoins en termes de gestion. La représentation de données massives et complexes est un champ de recherche très actif en informatique. Ainsi, de nouvelles technologies ont émergées, capables de traiter une grande variété de données et d'exécuter des applications à grande échelle sur des systèmes parallélisés, pouvant potentiellement impliquer des milliers de téraoctets de données. Elles sont regroupées sous les termes de NoSQL et NewSQL [70, 78, 149].

Récemment, les développements de nouvelles générations de base de données NoSQL, ont ouvert de nouvelles perspectives notamment en bioinformatique. Dans un premier temps, des applications dans le domaine génomique ont été développées avec CouchDB [137, 8], Cassandra [69] et MongoDB [187]. Puis les applications ont été élargies vers d'autres domaines comme la santé, le phénotype. Une étude comparative a également été faite dans le domaine de la sélection génomique et le breeding en agronomie [159]. Par ailleurs, le développement de framework d'analyses de données volumineuses en parallèle tels que Hadoop ou Spark, a donné lieu à des applications [184, 157, 205].

Concernant le développement de systèmes d'intégration, de nombreuses études ont montré que la représentation d'information sous forme de graphe était mieux adaptée pour gérer l'information biologique [85, 133]. Ainsi, nous constatons le développement d'un nombre croissant de base de données de graphes [84, 164]. Toutefois, ces applications utilisent une approche centralisée et fermée (i.e. les données ne sont pas facilement accessibles), à l'opposé des courants actuels qui encouragent l'open data (FAIR Data principles [230]) et l'interopérabilité des données [58, 123]. Elixir-Europe est une infrastructure Européenne impliquant les principaux instituts publics nationaux qui favorise le développement de services en bioinfomatique et favorise la dissémination

6. Le terme entrepôt en biologie ne correspond pas forcément aux entrepôts multi-dimensionnels

de l'information. Dans le domaine agronomique, des groupes de travail (RDA , DivSeek et PhenoHarmonIS) ont pour objectifs de promouvoir les bonnes pratiques de gestion et les standards d'échange de données.

La technologie Web Sémantique (SW) proposée par Tim Berners-Lee [18] offre une solution pour faciliter cette intégration et permettre l'interopérabilité entre les machines.

Au cours des dernières années, de nombreuses initiatives ont émergé dans la communauté biomédicale afin de fournir des environnements intégrés permettant de formuler des hypothèses scientifiques sur le rôles des gènes dans l'expression des phénotypes ou l'émergence de maladies. Parmi elles, citons BIO2RDF [16], OpenPHACTS [231] et EBI RDF [100]. Toutefois, il n'y a pas d'équivalent dans le domaine agronomique.

3.4 Représentation des données

3.4.1 Rappel sur le web de données

Les langages du Web de données

Le World Wide Web Consortium (W3C)[7] est mandaté de proposer différentes recommandations pour normaliser et rendre compatible les différentes technologies du Web. Imaginé par Tim Berners-Lee [18], le Web de données, encore nommé le Web sémantique, est une extension du Web. Soutenu par le W3C, le Web Sémantique s'est mis en place et développé dans de nombreux domaines pour composer le nuage d'information que nous connaissons aujourd'hui[8]. Le Web sémantique est représenté par une architecture multi-couches dont la pile est basée sur un identifiant unique de ressource (URI). Un URI est une série de caractères, utilisant le protocole HTTP pour décrire une ressource et ses composants, permettant ainsi l'identification des données sur le Web. Ainsi l'URI `http://purl.uniprot.org/uniprot/Q5K4R0.ttl` référence la protéine Q5K4R0 de la base de données Uniprot qui permet d'être directement accessible et interprétée par des machines.

Parmi les technologies utilisées pour exposer des données sur le Web de données RDF, RDFS, OWL et SPARQL sont les éléments importants[9]. **RDF (Resource Description Framework)**[10] est le standard de représentation utilisé pour intégrer des données issues de plusieurs sources. Il représente les informations sous la forme de triplets Subject-Predicate-Object. Ces triplets peuvent être combinés pour construire un grand réseau d'information (également connu sous le nom de graphe RDF), intégré à partir de différentes sources de données. **RDFS (Resource Description Framework Schema)**[11] fournit des éléments de base pour la définition de schémas destinés à structurer des ressources RDF. Il définit notamment la notion de classe *rdfs :Class* et sous-classe *rdfs :subClassOf* permettant de structurer les ressources RDF de manière hiérarchique. RDFS possède également la propriété *rdfs :Label* qui permet de nommer une ressource indépendamment de son URI. Il existe également d'autres langages tels que **OWL (Web Ontology Language)**[12] qui étend RDFS en offrant une meilleure expressivité pour structurer des ontologies. Grigoris Antoniou et Frank Van Harmelen apportent de plus amples informations sur le langage OWL [10].

7. `https://www.w3.org/`
8. `https://lod-cloud.net/`
9. Retrouvez plus de détails sur l'histoire du web dans le livre sur le Web Sémantique [71], de F. Gandon, C. Zucker, O. Corby, 2012, Ed. Dunod
10. `https://www.w3.org/2001/sw/wiki/RDF`
11. `https://www.w3.org/TR/rdf-schema/`
12. `https://www.w3.org/2001/sw/wiki/OWL`

Citons également, SKOS [13] (Simple Knowledge Organization System) adapté pour représenter des thésaurus et vocabulaires. Enfin, le langage de requête **SPARQL** [14] offre aux utilisateurs la flexibilité d'extraire et de manipuler les informations stockées sur plusieurs graphes RDF et même sur plusieurs bases de connaissances distribuées.

La section suivante donnera de plus amples détails sur la manière de représenter et d'interroger les données liées. Le rôle du Web sémantique dans cette masse de connaissances est de décrire les ressources pour favoriser leur exploitation. Reste qu'une grande partie des descriptions est écrite en langage naturel qui reste ambigu pour les machines ce qui amène une tentative de solution liée à l'usage d'ontologies (représentant divers concepts biologiques).

3.4.2 Exemples de représentation des données en biologie

Les données du Web sémantique sont représentées de façon hiérarchique sous forme de triplets RDF dans un multi-graphe orienté étiqueté.

Représentation des données

Comme illustré à la figure 3.7, les graphes sont constitués de sommets et d'arêtes représentant respectivement les ressources et les prédicats. Une fois que la ressource est identifiée par une URI, elle peut être sujette à une question (ou un prédicat) dont la réponse est l'objet qui est lui associé. L'objet peut être sous deux formes : soit une chaîne de caractères soit un URI. Il s'agit respectivement d'un littéral ou d'une ressource .

La figure 3.7 représente : la ressource (i.e. le sujet) par son URI, le prédicat *is_located_on* indique que ce sujet possède une position sur un chromosome, — la valeur de l'objet est *chromosome 1*.

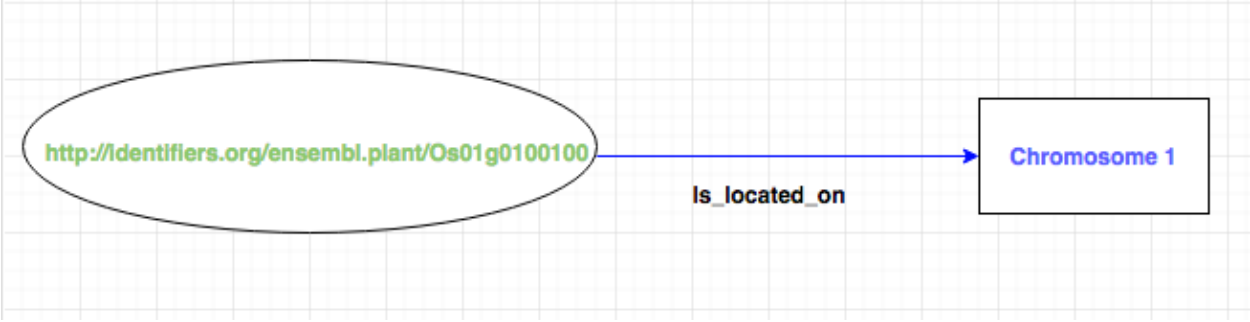


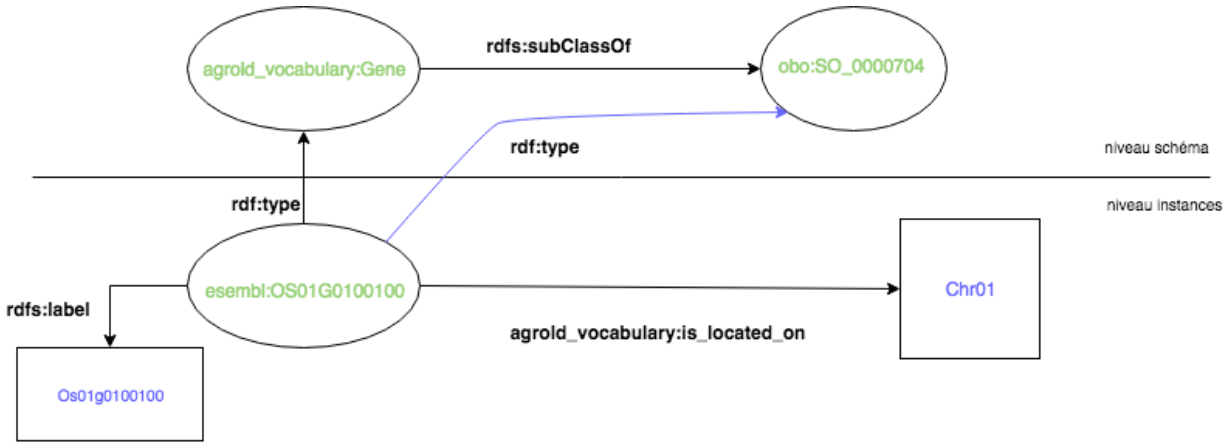
FIGURE 3.7 – Représentation d'un triplet RDF (sujet, prédicat, objet)

Utilisation de schémas sur les données

La figure 3.8 représente le triplet donné en exemple 3.7 dans le contexte du schéma RDF développé pour AgroLD (nommé agrold_vocabulary) [15]. La ressource identifiée par **ensembl :Os01g0100100** est décrite comme appartenant à la classe *Gene* du vocabulaire AgroLD par la relation *rdf :type* qui

13. https://www.w3.org/2009/08/skos-reference/skos.html
14. https://www.w3.org/TR/sparql11-query/
15. https://github.com/SouthGreenPlatform/AgroLD_ETL/tree/master/model

lui même est une sous-classe de la ressource **obo :SO_0000704** provenant de l'ontologie Sequence Ontology. Si on utilisait un mécanisme de raisonnement sur ces graphes et ontologies, on déduirait par transitivité que **ensembl :Os01g0100100** aurait aussi comme *rdf :type* **obo :SO_0000704**.




```
@base  <http://www.southgreen.fr/agrold/> .
@prefix rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix rdfs:<http://www.w3.org/2000/01/rdf-schema#> .
@prefix agrold_vocabulary:<http://www.southgreen.fr/agrold/vocabulary/> .
@prefix obo:<http://purl.obolibrary.org/obo/> .
@prefix ensembl:<http://identifiers.org/ensembl.plant/> .

ensembl:OS01G0100100
      rdf:type      agrold_vocabulary:Gene ;
      rdfs:label   "Os01g0100100"
      agrold_vocabulary:is_located_on    "Chr01"
```

FIGURE 3.8 – Représentation d'un schéma RDFS

Nous avons vu qu'il était possible d'insérer des informations sous la forme de triplets dans un graphe hiérarchisé afin de lier, représenter et publier les données sur le Web. Les ressources peuvent être décrites selon un vocabulaire déterminé tels que les noms de classes, les types de ressources et les types de relations entre elles. Cependant, RDFS connaît différentes limites dont nous ne citerons que deux exemples :

— la combinaison de classes : il n'est pas possible de montrer que la classe protéine a plusieurs familles,
— la disjonction : il n'est pas possible de dire que les oxydases et les réductases sont deux sous-classes disjointes.

De grands efforts ont été déployés depuis plus de 10 ans afin de structurer et partager les vocabulaires au sein de la communauté Biomédicale et des sciences de la vie. Parmi les initiatives les plus importantes citons MGED (Micro Array Gene Expression Data) [225] qui décrit les données d'expériences de puces à ADN mais surtout OBO (Open Biomedical Ontologies) [198, 77, 210] qui à travers un format standard (OBO), des outils (OBO Edit) et une plate-forme Web, centralise la majorité des ontologies développées dans le domaine biologique. Le projet a grandement contribué à la démocratisation et l'utilisation massive des vocabulaires contrôlés et des ontologies

dans ce domaine mené en premier lieu par Gene Ontology. Dés lors, de nombreuses plate-formes se sont développées afin de fournir un niveau de services important pour utiliser ces ontologies. Le National Center for Biomedical Ontologies (NCBO) développe et maintient Bioportal [158] une plateforme qui contient plus de 400 ontologies et terminologies biomédicales. Ontobee [161] permet de faciliter le partage d'ontologies, l'intégration et l'analyse des données ainsi que leur visualisation. L'Ontology Lookup Service (OLS) [45] fournit un accès aux ressources OBO utilisé par la plate-forme génomique de l'EBI (European Bioinformatics Institute).

Comme je l'ai déjà mentionné, les nouvelles technologies de production de données haut-débit en médecine et en biologie génèrent de grands volumes de données. En plus de ce phénomène qui touche également d'autres secteurs professionnels, s'ajoute l'hétérogénéité et la diversité et complexité des données qui sont les principaux problèmes abordés en bioinformatique. La recherche, le tri et l'accès aux données, leur interprétation et leurs annotations sont des tâches fastidieuses à réaliser pour un expert biologiste. Les technologies du Web sémantique offrent des solutions prometteuses pour ces domaines puisqu'elles visent à faire participer à la fois les utilisateurs et les machines [18].

Dans le Web sémantique, il existe différentes bases de connaissances qui sont regroupées en fonction du domaine d'expertise. **Gene Ontology** [16] **(GO)** [208] est une ontologie du domaine biologique représentant la connaissance sur la fonction des gènes et leur protéines associées. Créée il y a plus de 20 ans, d'abord comme un vocabulaire contrôlé et structuré, elle s'est rapidement imposée dans la communauté biologique comme un élément central pour l'annotation fonctionnelle des gènes. Aujourd'hui, c'est la ressource la plus utilisée et reliée aux autres ressources. Elle est composée de plus de 45 000 termes connectés par plus 134 000 relations. L'ontologie couvre 3 aspects distinct de l'annotation fonctionnelle des gènes : *la fonction moléculaire* (i.e. l'activité d'un produit de gène (protéine, enzyme, etc.) à un niveau moléculaire donné), *le composant cellulaire* (la localisation de l'activité du produit de gène au sein de la structure biologique), et *le processus biologique* (un niveau plus général du métabolisme biologique dans lequel la fonction biologique est associée). La base de connaissances comprend également des annotations, créées en reliant des produits de gènes (i.e. gènes, protéines, etc.) aux termes de l'ontologie. Chaque annotation comprend une information de provenance sur laquelle elle se fonde, comme une publication, qui est encodée avec des termes de l'ontologie Evidence and Conclusion Ontology (ECO) [36]. Par exemple, une annotation peut indiquer que « MSH2 Humain (un gène, HGNC :7325, également représenté par UniProtKB :P43246) est impliqué dans *GO :0006298 mismatch repair* (terme GO), basé sur une preuve *ECO :0000314* utilisée dans une annotation manuelle ». Formellement, cette annotation serait représentée dans la base de connaissances GO comme un triplet reliant le gène au terme GO en utilisant une relation spécifique : UniProtKB :P43246 *involved_in* GO :0006298. La base de connaissances GO contient plus de 7 millions d'annotations de gènes / produits de gènes sur plus de 3 200 espèces dont à peu prés 10% (750 000) sont validés par des données expérimentales issues de publications. Pour le reste, de nombreux logiciels ont été développés permettant de réaliser ces annotations de manière automatique, soit en se basant sur la séquence nucléotide du gène soit à partir de mot clés.

16. `http://geneontology.org/`

3.5 Extraction de connaissances biologiques

Le constat établi est que les ressources issues de bases de données restent limitées pour produire une connaissance suffisante et nécessaire pour formuler des hypothèses de recherche d'information sur les fonctions moléculaires des gènes et leurs rôles dans l'expression de phénotype [94]. Il existe des ressources annotées manuellement comme OryzaBase ou Qtaro (pour ne citer qu'un petit nombre) mais elles ne fournissent pas un contenu exhaustif de l'information et ont un long délai de mise à jour. L'extraction d'information à partir de texte biomédicaux n'est pas une tâche aisée. L'article de Champan *et al.*[32] en dresse un panorama à la fin des années 2000.

3.5.1 Méthodes d'extraction d'entités nommées

Un tâche importante dans le domaine bioinformatique concerne les méthodes d'extraction d'entités nommées (NER - Named Entity recognition). Les entités nommées sont par exemple les noms de gènes, protéines, d'espèces, de mutants ou de composés biochimiques. Par ailleurs, l'extraction de structures plus complexes telles que les relations et les évènements qui opèrent sur ces entités dépend de la capacité à détecter ces dernières. De fait, le domaine bénéficie d'une longue expérience en la matière. Les conférences **Biocreative** [17], **BioNLP** [18] en sont les vitrines depuis 2004.

Pour résoudre ce problème, plusieurs méthodes et outils de fouille de texte ont été développés et publiés dans la littérature. Ils sont répartis en quatre approches principales [15] : i) celles utilisant un dictionnaire de mots, ii) des méthodes à base de règles écrites manuellement, iii) d'autres utilisant des approches de machine learning, iv) enfin des approches combinant le machine learning et au moins une des deux précédentes.

Les méthodes basées sur des dictionnaires, l'une des approches bioinformatiques les plus fondamentales du NER, utilisent des listes complètes de termes afin d'identifier les occurrences d'entités dans le texte. Toutefois, compte tenu de l'évolution rapide et constante des découvertes scientifiques, il est difficile de maintenir à jour des listes de dictionnaires. De plus, l'abondance de synonymes (i.e. la même entité peut avoir deux noms différents) ou l'utilisation fréquente d'acronymes - MONOCULM (MOC) - complexifie la tâche d'identification et d'association à des entités existantes. Ces méthodes sont donc aujourd'hui associées à d'autres approches [87, 73].

Une autre approche consiste à définir des règles basées sur des modèles qui exploitent les caractéristiques orthographiques et lexicales des classes d'entités ciblées afin de les reconnaître (par exemple pour les protéines [66]). Parce qu'il nécessite une expertise humaine et beaucoup de travail pour créer de tels modèles, les systèmes ultérieurs ont essayé d'apprendre automatiquement de tels modèles à partir de données étiquetées [29, 37]. Des travaux plus récents sur la reconnaissance d'entités nommées utilisent des méthodes statistiques d'apprentissage automatique qui peuvent être combinées aux méthodes précédentes.

Au cours des dernières années, les deux méthodes précédentes ont été remplacées par des approches basées sur l'apprentissage automatique supervisé, en particulier les algorithmes de classification séquentielle, tels que les modèles de Markov cachés [172] et les CRF (*Conditional Random Fields*) [111]. Les CRF sont devenus le modèle standard *de facto* [191], étant la méthode de choix pour la quasi-totalité des outils ayant remporté des compétitions récentes de type NER, comme BioCreative IV [107] ou i2b2 [218] . Les outils NER populaires utilisant les CRF sont, par exemple,

17. Biocreative - `http://www.biocreative.org`
18. `http://2016.bionlp-st.org`

ABNER (*A Biomedical Named Entity Recognizer*) [190] et BANNER [119].
Les méthodes hybrides combinent des méthodes d'apprentissage automatique avec des techniques basées sur des dictionnaires ou des règles. Par exemple, ChemSpot [178] intègre les résultats d'un modèle CRF avec un module d'appariement de dictionnaire pour NER chimique.

3.5.2 Utilisation de méthodes d'apprentissage profond pour l'extraction d'entités nommées

Récemment, des méthodes de plongements lexicaux ou plongements de mots (« word embedding » en Anglais) ont permis d'améliorer les approches de fouille de texte. De manière générale, elles permettent une représentation de mots d'un dictionnaire sous la forme de vecteurs avec des nombres réels dans un espace à n-dimensions. Par exemple, le mot *homme* sera représenté par le vecteur ici à 2 dimensions [0,33 0,98]. En représentant tout les mots d'un dictionnaire selon la même méthode, il est facile d'imaginer que les mots ayant une proximité sémantique par exemple *homme* et *femme* auront des valeurs proches dans le même espace vectoriel [19]. De plus, ces représentations permettent d'envisager d'effectuer des opérations telles que *roi - homme + femme = reine* [144].

Dés lors, de nombreux modèles de représentation ont été développés afin de créer des plongements lexicaux à partir de mots. Parmi eux, citons les plus utilisés Word2Vec [145], Glove [165], ELMo [167] et BERT [53]. La principale différence entre ces modèles vient du fait que Word2vec et Glove ne prennent pas en compte l'ordre des mots dans la phrase (i.e. ils sont indépendant du contexte de la phrase ; dans cellule de prison et cellule sanguine, le mot cellule aura le même vecteur) alors que ELMo et BERT prennent en compte l'ordre des mots - mais avec deux approches différentes - (i.e. ils généreront des vecteurs différents pour les mêmes mots en fonction de leurs contexte ; dans l'exemple précédent le mot cellule aura deux vecteurs différents).

Les récents progrès des outils de fouille de textes biologiques ont été rendus possibles grâce aux avancées des techniques d'apprentissage profond utilisées dans le traitement du langage naturel (NLP). Par exemple, l'utilisation d'approches de réseaux de neurones combinés aux *Conditional Random Fields (CRF)* montrent des résultats bien meilleurs qu'avec les approches précédentes [186]. Notamment, les modèles de type LSTM-CRF *Long Short Term Memory (LSTM)* combinés avec CRF [113] offrent des résultats encourageants. Habibi *et al.* [80] ont adopté le modèle de Lample *et al.* [113] et utilisé des vecteurs de mots issus de plongements lexicaux (i.e. Word2Vec) comme vecteurs d'entrée dans un modèle bidirectionnel LSTM-CRF (Bi-LSTM-CRF). Cependant, ces méthodes nécessitent un volume de données important afin d'optimiser les phases d'entraînements [80]. Plus récemment, des améliorations ont été apportées à cette méthode en utilisant l'apprentissage par transfert [75], l'apprentissage en multi-couches [235] et l'apprentissage multitâches [224]. Finalement, les approches plus récentes utilisent des modèles de représentation plus sensibles aux contexte (i.e. ELMo et BERT). L'application DTranNER [89] utilise ELMo dans une architecture Bi-LSTM-CRF en améliorant l'étape de labellisation par CRF avec une structure d'apprentissage profond. L'application BioBERT [121] utilise le modèle BERT pour créer les vecteurs de mots contextualisés. Par ailleurs BioBERT a été entrainé sur des corpus biomédicaux.

19. Une application Tensorflow permet de voir les mots dans un espace vectoriel `http://projector.tensorflow.org/`

3.5.3 Méthodes d'extraction de relations

L'extraction de relations est un domaine de l'extraction d'information biomédicale et biologique qui a pris beaucoup d'importance au cours des deux dernières décennies. La majorité des travaux dans ce domaine sont concentrés sur les tâches d'extraction d'interactions entre molécules de médicaments (*Drug-Drug Interactions-DDI*), d'interactions entre protéines (*Protein-Protein Interactions-PPI*) ou les combinaisons des deux. De nombreux travaux ont été réalisés sur l'extraction de relations biomédicales en se concentrant sur les techniques basées sur des règles et l'apprentissage automatique. Ces techniques sont généralement classées en quatre groupes, à savoir les approches basées sur la co-occurrence, les modèles, les règles et l'apprentissage automatique [240].

L'approche la plus simple pour extraire les relations entre entités est la co-occurrence qui identifie les entités co-occurrentes dans une phrase, un résumé ou un document [72]. Les systèmes basés sur des motifs (*patterns*) s'appuient sur un ensemble de motifs pour extraire des relations ; ces motifs peuvent être définis aussi bien manuellement qu'automatiquement. Les modèles manuels sont définis par des experts du domaine, ce qui est un processus long et peu reproductible. Pour améliorer cette méthode, on utilise la génération automatique de modèles. La génération automatique de motifs peut utiliser le *bootstrapping* [222] ou générer directement à partir de corpus [129]. Dans les systèmes basés sur des règles, un ensemble de règles peut être construit pour extraire des relations [93]. Les systèmes basés sur des règles peuvent être définis de deux manières, manuellement et automatiquement.
Grâce à l'abondance de corpus annotés, les approches basées sur l'apprentissage machine, plus efficaces, se sont fortement développées. De nombreuses approches ont d'abord utilisé l'apprentissage supervisé, dans lequel les tâches d'extraction de relations été modélisées comme des problèmes de classification [26, 7]. Par la suite, des approches semi-supervisées [102] et non supervisées [171] ont été développées.

Plus récemment, l'utilisation des réseaux de neurones a pris également une place importante. *Zeng et al.* [238] ont d'abord utilisé les réseaux de neurones convolutionnels (CNN) pour capturer le mot ainsi que les informations de position pour l'extraction de relations. Leur modèle obtenait ainsi, une meilleure performance que les méthodes basées sur l'apprentissage machine. Un certain nombre d'efforts ont été faits récemment pour saisir des informations dans des séquences de texte plus longues en utilisant des modèles de réseau neuronal récurrent (RNN) [138] ou des modèles LSTM [128, 43]. *Zhou et al.* [241] et *Zheng et al.* [239] utilisent tous deux la structure LSTM pour modéliser des modèles de relations sur de longs segments de texte afin de capturer les informations sémantiques les plus importantes dans une phrase. Cependant les modèles LSTM ont des capacités de mémoire souvent limitées au mot et à son contexte. En se basant sur les avantages des réseaux de mémoire, certaines approches ont développé de nouveaux modèles de réseau de mémoire basés sur l'attention pour l'extraction de relations [130, 239]. Leurs modèles consistent en un réseau de mémoire au niveau du mot qui peut apprendre l'importance de chaque mot de contexte par rapport à la paire d'entités, et un réseau de mémoire au niveau de la relation qui peut capturer les dépendances entre les relations.
D'autres approches ont implémenté les modèles de plongements de mots avec contexte comme BERT pour représenter les relations [35].

Finalement, un autre type d'approche développée est de combiner l'information de relations extraites du texte analysé avec de l'information trouvée dans d'autres ressources comme des bases de données [243, 242]. Les données extraites de ressources externes sont intégrées sous forme de plongements de triplets (i.e. entité A relation entité B) dans un réseau de neurones. Elles sont

ensuite utilisées dans le modèle d'extraction de relations pour enrichir et filtrer l'information extraite.

Chapitre 4

Synthèse des activités de recherche et résultats obtenus

Comme énoncé précédemment, les progrès de la génomique et des outils de phénotypage à haut débit offrent une occasion unique de découvrir de nouveaux gènes. Cependant de nombreux challenges existent encore lorsqu'il s'agit d'exploiter les systèmes d'information actuels et le croisement de données massives et multi-échelles générées. Il s'agit notamment de traiter l'information sous-jacente en extrayant les connaissances incluses afin de découvrir rapidement des relations gène-phénotype et leur dépendance à l'environnement ce qui détermine le rendement des cultures dans des environnements divers.

Dans cette partie, en accord avec la problématique et les enjeux présentés dans le chapitre 3, nous présenterons les grands axes des propositions effectuées au cours de nos travaux de recherche concernant :

— l'interopérabilité des bases de données génomiques, l'intégration de ces données (section 4.1)

— les approches permettant la réécriture de requêtes entre différents systèmes d'information en effectuant l'enrichissement automatique de mapping BD-RDF (section 4.2) y compris dans le cas de bases NoSQL.

— la prise en compte des données massives (section 4.3) et l'impact sur la performance des requêtes ;

— l'enrichissement des connaissances grâce aux plateformes d'intégration de données sémantique (section 4.4)

4.1 Propositions d'approches décentralisées pour l'interopérabilité des bases de données agronomiques

Contexte Dans le contexte du projet ANR Génoplante, l'analyse fonctionnelle du génome du riz a été réalisée à partir de la création d'une collection de mutants de riz. L'étude comprenait le séquençage des gènes mutés ainsi que la caractérisation phénotypique des lignées de mutants sur plusieurs sites géographiques. J'ai eu la charge de développer le volet bio-informatique.
J'ai ainsi développé un workflow de détection et d'annotation des gènes disruptés (2004-02). J'ai également été impliqué, dans le développement de l'application OrygenesDB [57] [1], une application Web permettant de stocker des données relatives aux séquences générées lors du projet combinées aux séquences issues du séquençage du génome du riz (2006-01). Le principal objectif de mon travail a été la conception du système d'information dédié à la gestion des données phénotypiques et à leur enrichissement par des liens avec les autres ressources (génomique, transcriptomique, protéomique) décrivant la collection. Pour ce faire, j'ai développé OryzaTagLine [114] [2],

1. http://orygenesdb.cirad.fr
2. http://oryzatagline.cirad.fr

un système d'information permettant d'intégrer et centraliser ces différentes ressources afin de fournir un portail Web unique aux scientifiques (2008-01).

Problématique C'est au cours de ce projet, que je me suis engagé sur une thèse afin de lever de nombreux verrous liés à l'intégration de données dans le domaine agronomique. Les thématiques abordées concernaient (i) la formalisation de standards d'échange de données, de métadonnées et d'ontologies pour décrire et annoter les données, (ii) le développement d'approches permettant la communication de systèmes d'informations sur des réseaux distribués. L'objectif de mon travail était de proposer des solutions permettant aux scientifiques d'accéder de manière transparente aux informations issues de plusieurs sources de données (génomique, phénotypique, etc.). Pour cela, j'ai abordé le sujet en proposant deux approches : l'une basée sur une architecture de médiation (section 4.1.1) et l'autre basée sur une architecture orienté services (section 4.1.2).

4.1.1 Une architecture de médiation est elle adaptée dans ce contexte

Choix d'une approche générique de médiation

Dans un premier temps, nous avons fait le choix d'un système de médiation générique capable de répondre aux besoins du projet. Sur les conseils de mes encadrants de thèse, j'ai utilisé *Le Select* [135], après avoir assisté à une présentation de l'outil et suivi un tutoriel.

Successeur de DISCO [211], il avait une architecture de type médiateur/adaptateur qui utilisait un modèle pivot relationnel, afin d'intégrer de manière transparente les sources de données hétérogènes et distribuées. Le fait que *Le Select* utilisait le standard SQL comme langage de requête, lui permettait d'interagir avec un bon nombre de système d'information. Par ailleurs, de nombreux types de données pouvaient également être représentés de manière uniforme dans le modèle de données relationnel grâce à des adaptateurs pouvant être définis par l'utilisateur (e.g. structurés, semi-structurés, etc). Écrit en Java, le médiateur proposait également un accès uniforme à l'exécution de programmes intégrés (e.g. services, programmes) ainsi que la publication et le traitement des données issues de ces processus.

De manière générale, *Le Select* offrait des outils de transformation des données publiées et permettait d'y attacher une documentation structurée. D'un point de vue réseau, *Le Select* avait une architecture distribuée, ce qui veut dire qu'il n'existait pas de dépôt centralisé pour intégrer les données, ni de schéma global prédéfini. En effet plusieurs applications *Le Select* pouvaient coopérer pour fournir l'accès aux ressources. Publier par exemple, des systèmes d'information par l'intermédiaire de *Le Select* évitait de mettre à jour les données intégrées et maintenait leur autonomie vis à vis d'autres applications clientes. Ces avantages, nous voulions les mettre à profit dans le cadre d'un projet scientifique visant l'intégration de ressources de données végétales (2006-02).

Mise en œuvre de l'approche

Dans un premier temps, mes travaux ont consisté dans l'**intégration syntaxique** de plusieurs sources de données par la mise en place d'adaptateurs. Plusieurs instances ont été créées dont celles du CIRAD (incluant Orya Tag Line et OrygenesDB), une à l'IRD, une au CNRS (Univ. Perpignan) et une banque d'images au CIAT (Colombie). Dans de nombreux cas, il s'agissait d'instancier des librairies génériques proposées par le médiateur (base de données, fichiers structurés, exécution de programmes, etc.). Puis, j'ai développé des adaptateurs spécifiques aux formats de données spécifiques du domaines bioinformatiques tels que FASTA, EMBL, NCBI, UNIPROT ou

GFF. Par la suite, des métadonnées ont été générées et associées aux adaptateurs. Par exemple, des métadonnées ont été extraites automatiquement des systèmes de fichiers et des images pour instancier les adaptateurs de la banque d'image. Ces métadonnées étaient importantes pour que les médiateurs, communicant en réseau, identifient les sources intervenant dans une requête de médiateur.

Concernant l'**intégration sémantique**, *Le Select* ne proposait pas de mécanisme particulier pour détecter des correspondances (*mappings*) entre les éléments des sources. Seul des mécanismes de vues (identiques à ceux des bases de données relationnelles) existaient et pouvaient être utilisés à cette fin. Ainsi, j'ai pu développer des règles ACI (correspondance inter-schéma) au niveau des tables et des attributs pour chaque adaptateur. La traduction des éléments en ACI a été réalisée selon des règles établies par le document de spécification ODM (Métamodèle Définition Ontologie) [3]. Dans l'objectif de guider l'intégration de nouvelles sources de données, un schéma global sous la forme d'une vue *Le Select* a été développé sur la base des schémas exposés par les adaptateurs des instances. Les règles ACI ont été prises en compte dans sa construction. Enfin, afin de montrer l'intérêt d'une infrastructure de médiation distribuée dans ce contexte, une mise en œuvre de l'intégration sémantique a été illustrée à travers des exemples de recherche d'information biologique.

4.1.2 Comment la composition de services Web peut faciliter l'intégration

La deuxième approche proposait l'intégration des sources à travers l'enchaînement de services Web (SW) grâce à un environnement Web personnalisé (2008-02).

Choix d'une approche générique

La mise en œuvre de l'intégration de données par le biais de services Web nécessitait d'identifier différents composants tels qu'un protocole de communication entre services et un annuaire listant les services. Or dans la communauté bioinformatique, un tel système existait déjà : le framework BioMoby [229, 228]. BioMoby était un projet open source essentiellement orienté sur la découverte et l'exécution de SW biologiques. BioMOBY et son annuaire de services Web bioinformatiques utilisaient le protocole SOAP (Simple Object Access Protocol). Par le biais d'un annuaire central, l'application proposait aux fournisseurs d'enregistrer et de décrire leurs services en tenant compte d'un vocabulaire structuré. L'utilisation d'un tel vocabulaire pour décrire les SW permettait de faciliter la recherche et l'enchaînement des services. Toutefois BioMoby ne proposait pas d'outils permettant de gérer les conflits de noms, lors de l'enregistrement des services. Par ailleurs, il ne proposait pas non plus d'outils d'enchaînements de SW intégrés à son API.

Mise en œuvre de l'approche

Étant donné que BioMoby était très utilisé dans la communauté bioinformatique, de nombreuses applications et systèmes d'information y enregistraient leur services Web. Une des premières contributions a été de développer plusieurs services Web BioMoby pour les systèmes d'information Oryza Tag Line et OrygenesDB. Toutefois, BioMoby avait deux limitations importantes qui pouvaient freiner son développement : la gestion des enregistrements de services et l'orchestration de services. La première concernait l'enregistrement et la gestion des métadonnées au niveau du registre central. En effet, BioMOBY gérait mal la duplication des services, ou la présence de versions pour un même service. De plus, la manière de typer les services et leurs entrées/sorties

3. `https://www.omg.org/spec/ODM/About-ODM`

était très basique. Par exemple, elle ne permettait pas de créer une hiérarchie de classes ou de propriétés. Le système était limité dans sa capacité à proposer des possibilités d'enchaînements de services.

Ainsi, mes travaux ont porté sur la réalisation d'un système d'enchaînement de services BioMoby [57] (2007-01 et 2008-02). Ces travaux ont été réalisés avec un étudiant de master 2, Sébastien Fromentin, que j'ai co-encadré avec Gaetan Droc. Nous avons développé des méthodes pour orchestrer les SW que nous avions développés précédemment pour les systèmes d'information Oryza Tag Line et OrygenesDB afin de les combiner avec d'autres services BioMoby. Ces méthodes comprenaient des composants permettant de déterminer i) le type de données (ou d'objets) utilisé par le service, ii) rechercher les services compatibles, c'est-à-dire pouvant être parallélisés ou enchaînés en fonction du type de données utilisées, iii) déterminer et exécuter les étapes pouvant être exécutées en parallèle, iv) compiler et sérialiser les résultats dans divers formats (par exemple, proposer une fiche descriptive pour un gène dans un format HTML et CSV). Par ailleurs, une application Web a été développée afin de permettre l'utilisation de ces méthodes par un public de biologistes.

4.1.3 Conclusion

Les contributions développées au cours de ma thèse, ont permis de généraliser l'utilisation des standards d'échanges de données et des ontologies du domaine (au sein du CIRAD et de ses partenaires), ainsi que d'appliquer différentes approches d'intégration de données en agronomie. Même si l'approche de médiation présentait plusieurs avantages, comme par exemple, une accessibilité homogène aux systèmes d'informations et différents formats de fichiers, elle n'a pas été retenue pour la suite des travaux de recherche aux sein de notre unité. Cette approche de médiation avait en effet des contraintes plus importantes :

— des problèmes de sécurité d'accès aux bases de données. En effet, demander aux administrateurs les paramètres d'accès n'était jamais facile. L'alternative était de les aider à déployer les adaptateurs eux-mêmes, mais la configuration pouvait s'avérer complexe ;

— des problèmes de maintenance ou d'extension du schéma global et des règles de réécriture. Dans ce cas, l'ajout de nouvelles ressources nécessite une bonne connaissance du fonctionnement du médiateur, du schéma global et des règles ACI.

— il n'existait pas de projets similaires dans notre domaine sur lequel nous aurions pu compter et interagir avec.

L'approche orientée service Web présentait également des avantages et des inconvénients. Parmi les avantages nous pouvons citer :

— Le framework BioMOBY représentait une importante communauté de développeurs de service bioinformatique ;

— L'accès au système d'information est protégé et limité grâce aux principes des services Web ;

— L'accès au contenu du système d'information est limité. Il est décidé par le fournisseur ;

— Les entrées et sorties du service sont contraintes par des types ;

— Il est possible de rechercher des services existants par le registre central.

Parmi les inconvénients, ceux que nous avons mentionné dans la section précédente nous semblaient les plus importants. Néanmoins, nous avions déjà proposé des solutions pour certains d'entre-eux.

De manière générale, l'approche basée sur l'enchaînement de services Web nous semblait mieux adaptée au contexte de l'intégration de données pour notre communauté bioinformatique. De plus, le développement de service Web a également permis d'accroître les fonctionnalités des applications OryzaTagLine et OrygenesDB et leur interopérabilité avec d'autres systèmes existants.

Sélection de références

— (2004-02) Sallaud C., Gay C., **Larmande P.**, Bès M., Piffanelli P., Piégu B., Droc G., Regad F., Bourgeois E., Meynard D., Périn C., Sabau X., Ghesquière A., Delseny M., Glaszmann J.C., Guiderdoni, E. (2004) High throughput T-DNA insertion mutagenesis in rice : A first step towards in silico reverse genetics. *Plant J.* 2004 Aug ; 39(3) :450-64. Impact Factor : 5.468

— (2006-01) Droc G, Ruiz M, **Larmande P**, Pereira A, Piffanelli P, Morel JB, Dievart A, Courtois B, Guiderdoni E, Périn C. OryGenesDB : a database for rice reverse genetics. *Nucleic Acids Res.* 2006 Jan 1 ; 34 (Database issue) :D736-40. Impact factor : 9.202

— (2006-02) **Larmande P**, Tranchant-Dubreuil C, Regnier L, Mougenot I, Libourel T. Integration of Data Sources for Plant Genomics. *ICEIS* (1) 2006 : 314-318

— (2007-01) Larmande P. A personalized integrated system for rice functional genomic. 2007, Poster, *NETTAB*, Pise, Italie.

— (2008-01) **Larmande P**, Gay C, Lorieux M, Périn C, Bouniol M, Droc G, Sallaud C, Perez P, Barnola I, Biderre-Petit C, Martin J, Morel JB, Johnson AA, Bourgis F, Ghesquière, A, Ruiz M, Courtois B, Guiderdoni E. Oryza Tag Line, a phenotypic mutant database for the Genoplante rice insertion line library. *Nucleic Acids Res.* 2008 Jan ; 36 (Database issue) :D1022-7. Impact factor : 9.202

— (2008-02) Droc G, Périn C, Fromentin S, **Larmande P**. OryGenesDB 2008 update : database interoperability for functional genomics of rice. *Nucleic Acids Res.* 2009 Jan ; 37 (Database issue) :D992-5. Impact factor : 9.202

4.2 Enrichissement automatique de mappings BD-RDF pour faciliter la réécriture de requêtes

Contexte Ayant rejoint l'équipe intégration de données de M. Ruiz, j'ai été impliqué dans le projet international Generation Challenge Programme (GCP), l'un des cinq Challenge Programme établis par le Consultative Group on International Agricultural of Research (CGIAR).

Une plate-forme d'intégration de données, nommée GCP Pantheon, avait été développée afin de permettre à des clients logiciels d'interroger de manière transparente tout type de données produites dans le cadre du programme GCP. Cette plate-forme combinait une approche de médiation LAV (Local As View) et une approche sémantique, permettant aux partenaires de gérer leurs données localement puis de les rendre accessibles facilement en accord avec un modèle conçu spécifiquement pour le projet, le GCP Domain Model [25, 221]. Le modèle GCP a également été utilisé pour implémenter le framework BioMoby qui s'était rapidement imposé comme protocole standard d'échange de données et de services dans le domaine bioinformatique.

Problématique La principale limitation de la plate-forme GCP Pantheon était due au « mapping » manuel des schémas des sources locales sur le schéma global du GCP Domain Model. Afin de lever cette limitation, j'ai co-encadré une thèse (thèse de Julien Wollbrett) sur les méthodes de création automatique d'adaptateurs, facilitant ainsi l'intégration sémantique des bases de données

relationnelles biologiques sur la plateforme GCP. A cet effet, la thèse exploitait l'utilisation de différentes ontologies de domaines (génomique, phénotypiques, etc.) permettant l'établissement à la fois des règles de correspondance et d'interprétation, nécessaires à l'intégration automatisée.

Le point commun de l'intégration de données et des technologies du Web sémantique est de traiter et exploiter la connaissance liée à l'hétérogénéité sémantique de sources de données interconnectées. Le Web sémantique facilite la représentation de la sémantique des données et peut ainsi être utilisé pour faciliter l'interopérabilité ou l'intégration de données [9, 97]. Parmi les technologies utilisées pour exposer des données sur le Web de données RDF, RDFS, OWL et SPARQL sont les éléments importants.

4.2.1 État de l'art des approches de mappings BD-RDF

RDF (Resource Description Framework) est largement utilisé pour intégrer des données issues de plusieurs sources. Ceci est dû au cadre qu'il fournit pour décrire, une ressource et ses relations, sous la forme de triplets Subject-Predicate-Object. Ces triplets peuvent être combinés pour construire un grand réseau d'informations (également connu sous le nom de graphe RDF), intégré à partir de différentes sources de données. La transformation de base de données relationnelles (BDR) en RDF est confrontée à la problématique de mise en correspondance (*mapping*) entre des schémas de BDR et une représentation sous forme de graphe RDF. De nombreuses approches de mapping entre BDR et RDF ont été proposées ces dernières années afin de répondre à plusieurs motivations. La production de logiciels implémentant ces diverses approches fut toutefois marquée par la plateforme D2RQ [21, 22] et par la spécification récente d'un langage de mapping nommé R2RML [200] dont la recommandation est apparue après nos travaux. Nous renvoyons les lecteurs vers une synthèse exhaustive des approches et outils de mapping (Michel *et al*) [142].

Peu d'outils disponibles au début du projet (2010) utilisaient des standards du Web sémantique, que ce soit pour la vue du schéma de la BDR (RDF, XML), ou pour le langage de requête (SPARQL), et permettaient de faire correspondre une base de données avec plusieurs ontologies. Dans notre approche, nous avons choisi d'utiliser D2RQ. Il s'agit d'une plate-forme de publication de BDR sur le Web utilisant les standards du Web Sémantique et permettant de traiter une base de données relationnelle comme un graphe virtuel RDF. Dans ce graphe, un élément du schéma est représenté par un nœud et une relation par un arc orienté. Il est possible de créer ce graphe virtuel RDF en exportant uniquement le schéma de la BDR. Nous parlerons alors de vue RDF.

La plate-forme D2RQ est composée de trois éléments principaux : i) Le langage déclaratif de mapping D2RQ, utilisé pour créer la vue RDF de la BDR et permettant de décrire les relations entre des ontologies et un schéma de BDR ; ii) Le moteur D2RQ permettant de créer automatiquement une vue RDF et de ré-écrire une requête SPARQL en une requête SQL interrogeant directement la BDR ; iii) Le Serveur HTTP D2R permettant d'interroger les bases de données relationnelles via le Web.

L'approche que nous décrivons ci-dessous détourne D2RQ pour homogénéiser des schémas hétérogènes et automatiser la création de requêtes sur des BDR distribuées. Pour cela nous avons enrichi sémantiquement et de manière automatique la vue RDF du schéma de BDR créée par D2RQ (décrit en section 4.2.2). Nous avons ensuite travaillé sur la formulation de requêtes se basant sur les vues RDF ainsi générées en développant un algorithme de recherche de plus court chemin dans les graphes RDF [233] capable de prendre en compte les particularités des schémas relationnels (décrit en section 4.2.3). Finalement nous proposons le framework BioSemantic, une

approche flexible, générique et automatisée en nous appuyant sur des standards du Web Sémantique et des Services Web (2013-01).

4.2.2 Enrichissement sémantique des mappings BDR-RDF

Nous souhaitons utiliser le langage D2RQ pour parcourir notre vue RDF et ainsi indirectement parcourir notre schéma de BDR. Toutefois, D2RQ n'ayant pas été implémenté pour cette utilisation, le langage D2RQ n'est pas suffisamment expressif pour définir toutes les relations que nous souhaiterions.

En effet, une des spécificités d'un schéma conceptuel de BD (exprimé en formalisme Entité Association par exemple comme le représente la figure 4.1) par rapport à un simple graphe RDF, est la présence de relations bien définies entre les schémas des tables.

Les relations concernées sont :

— **La relation d'agrégation :** par défaut, une association exprime une relation à couplage faible. Les classes associées restent relativement indépendantes l'une de l'autre [168]. L'agrégation est une forme particulière d'association qui exprime un couplage plus fort entre classes. Elle permet d'exprimer des relations de type maître/esclaves et représente des connexions bidirectionnelles dissymétriques.

— **La relation de composition** : il s'agit d'une forme d'agrégation avec couplage plus important entre les classes. Cette composition indique que la destruction du composite entraîne automatiquement la destruction des composants agrégés.

— **La relation dite "d'héritage"** : la généralisation et la spécialisation sont des points de vue portés sur les hiérarchies de classes. Une classe A est une spécialisation d'une classe B si chaque instance de A est une instance de B et si chaque instance de B est associée à au plus une instance de A.

D2RQ ne permet pas de prendre en compte ces spécificités or ce sont ces relations qui enrichiront la vue RDF.
Pour savoir comment prendre en compte ces types de relation, il faut s'intéresser aux règles de conversion de ce type de relation du modèle conceptuel au modèle relationnel.

Passage au modèle relationnel

Le passage d'un modèle conceptuel E-A en schéma de base de données obéit aux règles suivantes :

— Toute classe d'entités donne lieu à un schéma de table comportant l'ensemble des attributs de la classe ; chaque schéma a une clé primaire.

— Les relations binaires classiques (d'arité 2), agrégations, compositions ne donnent naissance à un schéma de table que si les cardinalités exprimées à leur deux extrémités sont de type (0,n) ; ce schéma comporte comme attribut (clé primaire) la concaténation des deux clés primaires des schémas correspondant aux classes reliées. Lorsqu'une des cardinalités exprimée est de type (0,1), il suffit de créer dans le schéma de la table correspondant à la classe non reliée à cette cardinalité, une clé étrangère référençant la clé primaire du schéma correspondant à l'autre classe.

— La règle concernant la conversion des classes liées par une relation d'héritage peut s'effectuer de trois façons différentes [59] :

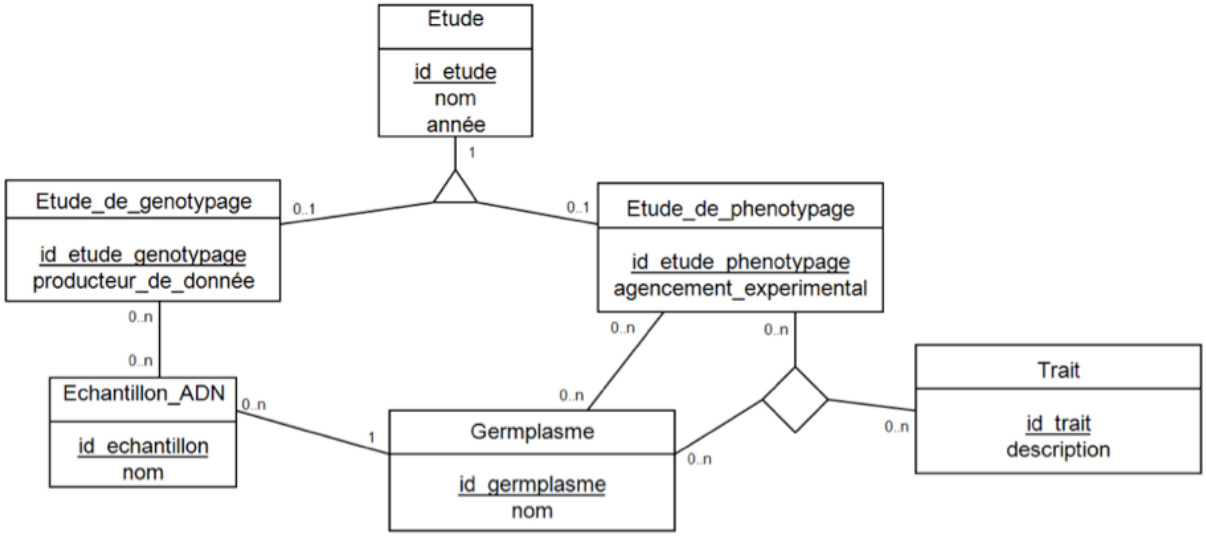


FIGURE 4.1 – Schéma conceptuel E-A où figure une relation d'héritage.
Les classes d'entités Etude_de_genotypage et Etude_de_phenotypage sont des spécialisations de la classe d'entités Etude

— soit en ne créant que le schéma de la table correspondant à la classe la plus générale et en faisant remonter les attributs de la sous-classe plus spécifique dans la description de celui-ci (on parle d'aplatissement vers le haut).
— soit en ne créant que le schéma correspondant à la classe la plus spécifique et en faisant descendre les attributs de la super-classe dans la description de celui-ci (on parle d'aplatissement vers le bas).
— soit, contrairement aux deux règles précédentes, on souhaite respecter la hiérarchie c'est-à-dire les niveaux général/spécifique, en créant les schémas des deux tables correspondant aux deux classes liées par la relation d'héritage avec création d'une clé étrangère supplémentaire dans le schéma de la classe spécialisée correspondant à la clé primaire du schéma de classe générale.

Détection automatique des relations.

Nous aurons besoin ultérieurement, lors de la création de requêtes, de détecter tous les types de chemins que nous souhaiterions combiner pour répondre à celles-ci. La détection automatique des relations du modèle conceptuel est alors intéressante. Pour la relation d'héritage, cette détection utilise la particularité des contraintes d'intégrité définies entre les schémas des tables des classes généralistes et ceux des tables des classes spécifiques. En effet, pour qu'une table soit une spécialisation d'une autre table, elle doit contenir pour seule clé étrangère la clé primaire de la table généraliste. La détection automatique de relations d'héritage pour la transformation d'un schéma vers une ontologie a été décrite dans [209]. Elle est également utilisée pour typer les relations d'héritage de l'outil DB2OWL [46]. Dans un article plus récent, les auteurs démontrent que ce type de transformation n'est possible que dans le cas où les BDR respectent la troisième forme normale [189]. La détection automatique de ce genre de relation d'héritage est alors rendue possible en utilisant la règle suivante :

```
Subclass(r,s) <- Rel(r)^Rel(s)^PK(x,r)^FK(x,r,_,s)
```

Cette règle, extraite de l'article de Sequeda *et al* [189], indique qu'une classe d'entités r correspondant au schéma relationnel Rel(r) est une sous-classe d'une classe d'entités correspondant au schéma relationnel Rel(s), si la clé primaire de Rel(r) est une clé étrangère de s.

L'utilisation de cette règle rend automatique la détection de toutes les relations d'héritage non aplaties.

La détection concerne également les relations d'agrégation ou de composition. L'implémentation de cette règle nous permettra ultérieurement donc de détecter tous les types de chemins que nous souhaiterions combiner lors de la création de nos requêtes.

Vues RDF enrichies

Dans notre cas, la détection de relation d'héritage implique la création de deux nouveaux triplets dans la vue RDF enrichie comme dans l'exemple ci-dessous correspondant au schéma 4.2.

```
etude_de_genotypage     rdfs:subClassOf         etude
etude_de_phenotypage    rdfs:subClassOf         etude
```

La prise en compte des relations précédentes permet de compléter l'enrichissement de la vue RDF. Pour cela, nous allons dans un premier temps prendre en compte les tables correspondant aux associations Les tables d'associations possédant ou non des attributs sont annotées avec la propriété *dr :associatedTo*. Un triplet contenant ce prédicat sera ajouté pour chaque table associée. Le sujet de ce triplet correspondra à la table d'association et l'objet à la table associée. L'arité d'une table d'association est annotée avec la propriété *dr :arity*. Ce typage est réalisé automatiquement, sous la forme de triplets, lors de la création de la vue RDF.

Nous obtenons une vue RDF dont la représentation sous forme de graphe est présentée dans la Figure 4.2.

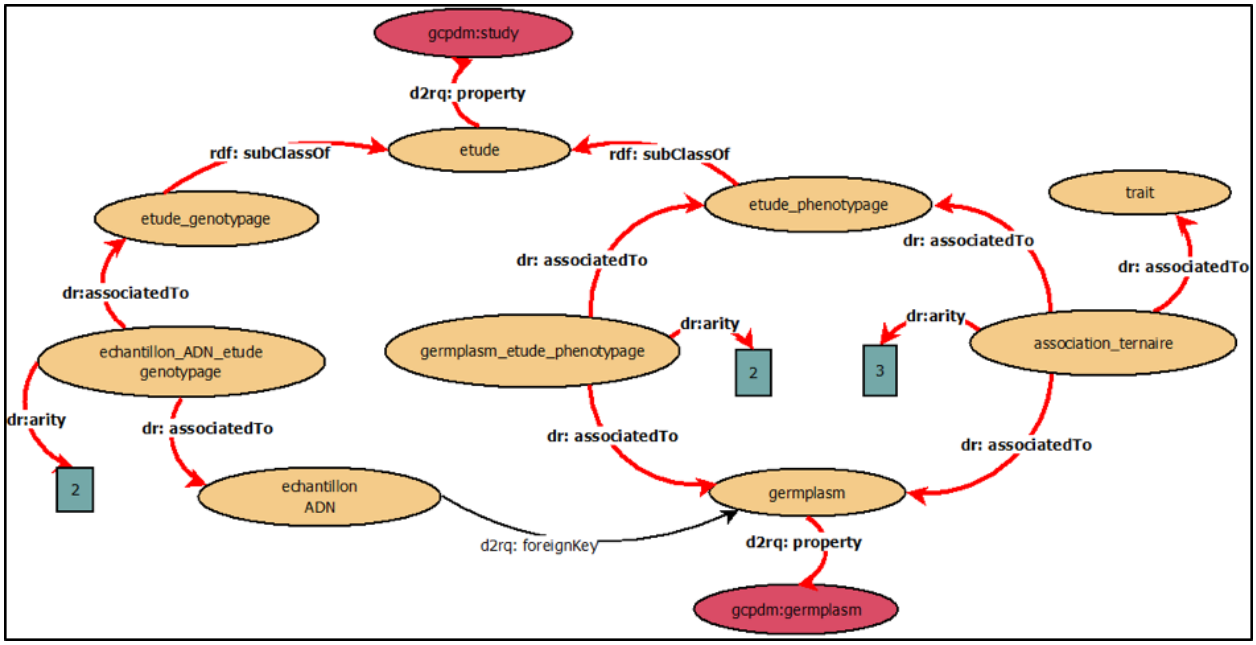


FIGURE 4.2 – Graphe représentant la vue RDF enrichie du schéma de la base de données utilisée.

Dans ce graphe, seul les nœuds représentant des tables sont présents, ces nœuds sont de couleur orange. Les nœuds rouges représentent les annotations sémantiques, ajoutées manuellement, réalisées sur une colonne de la table associée. Les arcs noirs représentent des propriétés présentes

d'origine dans la vue RDF, et les arcs rouges représentent les arcs rajoutés automatiquement par notre approche. Les nœuds bleus représentent la valeur associée à l'arité d'une table d'association, qui est détectée automatiquement.

4.2.3 Méthode de réécriture de requêtes SPARQL à partir de contraintes de schémas

Pour générer des requêtes SPARQL à partir de la vue RDF enrichie, nous souhaitions utiliser un algorithme de plus court chemin. Pour créer une requête SPARQL, nous utilisons les annotations sémantiques ajoutées manuellement dans la vue RDF du schéma de BDR concerné. Dans l'exemple de la Figure 4.2, nous allons sélectionner les annotations sémantiques *gcpdm :etude* et *gcpdm :germplasm* pour créer automatiquement une requête renvoyant tous les germplasms d'une étude donnée.

Lors de la recherche du plus court chemin, l'algorithme va détecter une relation de spécialisation entre la table *Etude* et les tables *Etude_genotypage* et *Etude_phenotypage* grâce à l'enrichissement sémantique avec les balises *rdf :subclassOf*. Cette information va être prise en compte et le plus court chemin renvoyé sera donc l'agrégation des plus courts chemins passant par ces 2 tables spécialisées (flèches rouges de l'étape A de la Figure 4.3). Lors du passage par la table *Etude_phenotypage*, l'algorithme a la possibilité de trouver 2 chemins passant par le même nombre de nœuds. Le premier chemin passe par la table d'association binaire *germplasm_etude_phenotypage* et le deuxième chemin par la table d'association ternaire appelée ici *association_ternaire*. L'enrichissement sémantique avec les balises *dr :associatedTo* permet à l'algorithme de détecter l'arité de ces tables d'association et de choisir de passer par celle ayant l'arité la plus petite. Dans l'exemple de l'étape B de la Figure 4.3, l'algorithme passera par la flèche rouge de gauche et ne parcoura pas la portion de graphe passant par la flèche rouge droite. Le plus court chemin final renvoyé par notre algorithme, permettant de créer automatiquement une requête pertinente, est représenté dans l'étape C de la Figure 4.3.

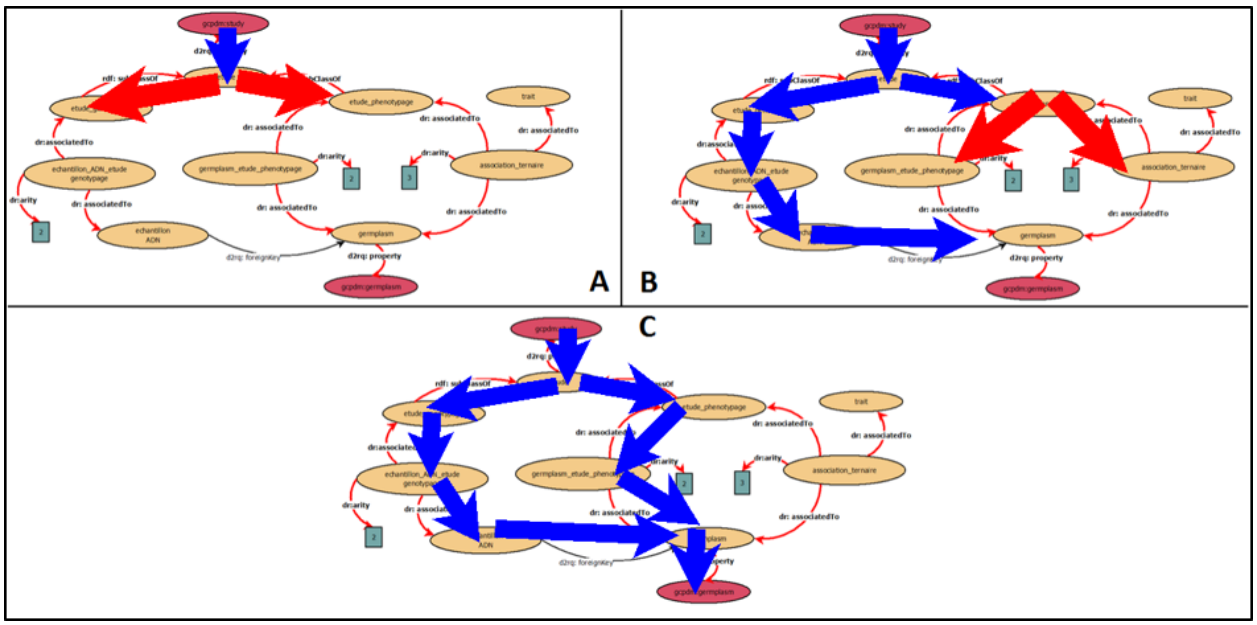


FIGURE 4.3 – Utilisation de l'enrichissement sémantique dans le parcours de graphe.

4.2.4 Évaluation d'approches de mapping NoSQL-RDF

Contexte Plus récemment, ayant en perspective d'extraire de la connaissance de la masse d'information biologique, nous avons commencé à explorer les possibilités que peuvent offrir les approches de mapping entre le Web sémantique et les bases de données NoSQL (e.g., MongoDB). Dans le cadre du projet BIOeSAI obtenu par financement incitatif interne IRD, un système d'information a été développé pour stocker des expérimentations en utilisant MongoDB (2015-01). Ces expérimentations requièrent la manipulation d'un volume important de données qui de fait sont de nature hétérogènes et stockées sous des formes différentes (fichier Excel, texte structuré ou semi-structuré, images, etc.). Ce volume et cette diversité de données peuvent rendre leur exploitation par les chercheurs difficile et non optimale. Dans ce contexte, un système d'intégration et d'indexation générique a été développé afin de pouvoir naviguer, partager et annoter ces données dans le but de les exploiter au mieux.

Problématique Ce système incluait également la gestion des métadonnées et des annotations ajoutées par les utilisateurs. Toutefois, la méthode mise en place ne permettait pas de détecter des relations explicites/implicites entre les données gérées par le système. Par exemple, il n'était pas possible de déduire qu'une région géographique (localisation GPS ou textuelle) est incluse dans une région plus large afin d'agréger des résultats. Ou encore, il était impossible de propager une information vue comme implicite, par exemple, « une maladie affectant une plante affectera tous ses tissus. »

Au cours de son stage de Master 2, Luyen Le Ngoc évalua plusieurs approches pour mettre en correspondance le schéma du système d'information MongoDB, une base NoSQL de type document structurée en JSON, à un shéma RDF annoté avec des ontologies biologiques [132] (2016-02). Il aborda notamment les approches de matérialisation de données en triplets RDF avec xR2RML [143] et de ré-écriture de requêtes avec les applications Ontop [179], xR2RML [141] et AllegroGraph. Puis dans une seconde optique de matérialisation, il évalua le SGBD de graphe NEO4J à partir d'import de données JSON. Enfin, il évalua également, l'utilisation de MongoDB avec des documents JSON-LD comme source de stochage et l'API RDF Jena pour la gestion des triplets RDF.

La solution retenue fut l'approche xR2RML avec une matérialisation en RDF pour des bases de petites tailles et une ré-écriture pour des bases plus importantes. Toutefois, cette dernière approche n'a pas été évaluée faute de temps.

Afin de répondre à la question de passage à l'échelle pour la gestion de triplets RDF, nous avons également évalué plusieurs triple-stores : Sesame, 4Store, Virtuoso, Jena Fuseki, StarDog, AllegroGraph et GraphDB.
L'évaluation porta sur différents critères à savoir :

— le chargement de données ;

— la recherche d'information avec projection, filtre, tri, union ;

— la recherche d'information avec plusieurs type d'inférences.

Une architecture générique a été développée afin de tester les différentes opérations sur les triple-stores. Une couche médiatrice logicielle orchestrait les différentes tâches à tester. Les tests ont été réalisés sur des données réelles produites par le projet Phénome et stockées dans la base de données PHIS (INRA) [154]. Elles comportaient à la fois des données textuelles et des images avec méta-données associées. Une ontologie développée par l'équipe de Pascal Neveu (Phis) a été utilisée pour effectuer des requêtes d'inférences sur les données transformées en RDF.

Parmi les solutions commerciales, StarDog a obtenu de très bon résultats sur l'ensemble des tests. Virtuoso édition libre, a obtenu de bons résultats pour les logiciels libres. Ces résultats publiés (2016-02) [132], nous ont conforté dans l'utilisation de Virtuoso pour la suite de nos travaux.

4.2.5 Conclusion

BioSemantic est un outil de médiation de données utilisant les technologies du Web sémantique et alliant à la fois les principes de médiateur/adaptateur et de services Web à une ou plusieurs sources de données. Il présente les avantages de 1) s'appuyer sur un schéma global pour construire des mappings BDR-RDF, 2) utiliser ces mappings pour ré-écrire les requêtes entre le médiateur et les BDR, 3) créer et encapsuler les requêtes dans des services Web. Toutefois, BioSemantic présente des limitations. Un premier élément limitant est son étape semi-automatique d'enrichissement sémantique des mappings. Un second élément est sa limitation d'utilisation aux BDR. Enfin, la création de services Web est basée sur le protocole SOAP et le Framework BioMOBY, aujourd'hui moins utilisés. De ce fait, dans la suite de nos travaux nous avons préféré opter vers de nouvelles solutions qui élargissaient les possibilités d'utilisation et réduisaient les limitations mentionnées.

Les travaux récents que nous avons menés sur les systèmes d'information NoSQL nous ont également fait tester des solutions existantes pour les SGBDR. Il parait évident que les choix de transformer entièrement ou partiellement - en utilisant la ré-écriture de requête - une source de données en RDF doivent tenir compte de certains critères comme la fréquence des mises à jours, le volume des données, la complexité des requêtes et le raisonnement qui peut être effectué sur les données. En se basant sur les travaux menés en section 4.2.4 nous préconisons une matérialisation en RDF pour des bases de petite taille ou faible fréquence de mises à jour. Par exemple, RML [54], MorphRDB [4] et xR2RML sont des outils adaptés et maintenus. Pour des systèmes avec d'importants volumes de données mises à jours fréquemment nous préconisons une ré-écriture de requêtes. Les outils Ontop et xR2RML permettent de mettre en œuvre cette méthode, toutefois nous n'avons pas eu le temps de les évaluer complètement.

Sélection de références

— (2013-01) Wollbrett J, **Larmande P**, De Lamotte F, Ruiz M. Clever creation of rich SPARQL queries from annotated relational schema : application to Semantic Web Service creation. *BMC Bioinformatics*. 2013. Impact Factor : 2.435

— (2016-02) Le Ngoc L, Tireau A, Venkatesan A, Neveu P, **Larmande P**. Development of a knowledge system for Big Data : Case study to phenotyping data. *Int. Conf. Web Intell. Min. Semant. Proceedings ACM WIMS '16*. 2016. Nimes (France)

— (2015-01) Le Ngoc L., Jouannic S. and **Larmande P**. Développement d'un outil générique d'indexation pour optimiser l'exploitation de données biologiques. *Poster aux Journées ouvertes pour la Biologie, l'informatique et les Mathematiques JOBIM* 2015. Clermont-Ferrant (France)

4.3 Passage à l'échelle dans la gestion des données génomiques

Contexte et problématique Les enjeux du stockage et traitement des données génomiques sont au cœur des problématiques de l'unité DIADE. L'équipe RICE coordonne le projet IRIGIN (International RIce Genomic Initiative) dont l'objectif est de réaliser le re-séquençage de centaines de variétés de riz et le génotypage par séquençage de milliers de lignées de riz, avec l'équivalent de 18 000 génomes de riz en termes de volume de données (90 TeraBytes). Bien que les biologistes

4. https://github.com/oeg-upm/morph-rdb

manipulent encore leurs données sur leur poste de travail à l'aide de logiciels de type tableur, aujourd'hui, leur limite d'utilisation est atteinte face à des données massives et complexes. Il en résulte que les alternatives souvent proposées jusqu'alors, passent par l'utilisation de traitements automatiques exécutables en lignes de commandes ou ont recours à des SGBD de type relationnels qui passent difficilement à l'échelle dans certains cas. Avec mes collègues, nous souhaitions lever ces verrous en proposant une nouvelle approche utilisant les SGBD NoSQL.

4.3.1 Comment combiner le stockage massif et les performances de requêtes

Depuis 2013, je suis impliqué dans le développement d'un logiciel facilitant cette tâche. Gigwa est une application qui utilise la technologie de base de données NoSQL (MongoDB) afin de gérer le passage à l'échelle pour le stockage et l'analyse des données de variations génomiques (typiquement issus de fichiers de format VCF, le standard de représentation des variations génomiques), et d'offrir une interface Web permettant d'y appliquer des filtres. Ce système permet alors de naviguer dans les résultats et de ré-exporter des sous-jeux de données dans divers formats de données et de visualiser les variations dans leur contexte génomique.

La contribution novatrice dans ce projet réside dans le modèle de stockage de données, que nous avons défini et optimisé pour ce type de données. De plus, le modèle tire avantage de la flexibilité d'extension du SGBD et permet d'utiliser l'application sur un ordinateur de bureau comme sur un cluster de calcul en distribuant les données sur plusieurs nœuds. Bien entendu, les performances tiennent compte du volume de données stockées et des ressources allouées, mais nous obtenons des résultats encourageants par rapport aux autres applications importantes dans ce domaine. Un article effectuant le comparatif et décrivant l'application a été publié en 2016 (2016-01) [187].

Récemment, nous avons développé une nouvelle version de Gigwa (V2) [188] (2019-01). Elle comprend notamment une API de services Web REST comme alternative à son interface Web. Cette API REST implémente et étend les recommandations du GA4GH Data Working Group [5] et de la Breeding API [6]. Par ailleurs, nous avons contribué au développement de la Breeding API [1] (2019-03) afin qu'elle soit en accord avec le modèle de données que nous développions. Cela permettra d'accroître l'interopérabilité de Gigwa avec d'autres systèmes utilisés dans la communauté bioinformatique tels que Galaxy [74, 76], FlapJack [146], SniPlay [51] ou Toggle [148]. De plus, afin d'accroître la flexibilité d'utilisation de Gigwa, nous avons développé plusieurs adaptateurs à des formats d'imports et d'exports de données standardisés dans le domaine bioinformatique (Plink, BED, Darwin, GVCF, GFF, etc.). Contrairement à des logiciels bioinformatique traditionnels, ces adaptateurs tirent partie du SGBD sous-jacent pour recalculer à la volée les données à exporter en fonction des paramètres modifiés. Par ailleurs, nous avons pu amélioré la vitesse d'exécution des requêtes, par rapport à la version 1, en optimisant les différentes étapes des requêtes. Dans notre dernier article [188], nous avons effectué un comparatif des vitesses d'exécution entre les versions 1 et 2 sur plusieurs types de requêtes ainsi qu'avec d'autres logiciels.

Dans un autre contexte, nous avons également participé à un comparatif de systèmes d'informations permettant de gérer les données génomiques. Nous avons sélectionné et évalué sept systèmes de stockage de données open source populaires (PostgreSQL, MonetDB, MariaDB, HDF5, Elasticsearch, Spark et MongoDB). MongoDB obtient de bons résultats dans certain type de requêtes mais est distancé par des systèmes hybrides qui tirent partie de leur flexibilité. Les résultats du comparatifs ont été publiés dans un journal bioinformatique spécialisé (Database) [160]. Toutefois, ces tests ne tenaient pas compte des améliorations faites sur la version 2 de Gigwa.

5. `http://ga4gh.org`
6. `https://brapi.org`

4.3.2 Conclusion

Comme nous avons pu le voir à travers les différents comparatifs que nous avons publiés, Gigwa est une application qui obtient de bons résultats parmi un ensemble d'applications. De par sa convivialité et son accessibilité, son interface utilisateurs est un avantage certain comparé aux logiciels en ligne de commande. Toutefois, c'est au détriment de sa vitesse d'exécution qui peut être inférieure dans certains cas. La gestion de très gros volumes de données, pour de très grands projets de séquençage >1To, peut être également un facteur limitant à son utilisation. Nous y travaillons. Il est fort possible qu'à l'avenir nous nous orientons vers différentes stratégies de déploiement, c'est à dire, optimisées pour l'ordinateur de bureau et pour de grosses infrastructures de calcul.

Sélection de références

— (2016-01) Sempéré G, Philippe F, Dereeper A, Ruiz M, Sarah G, **Larmande P**. Gigwa—Genotype investigator for genome- wide analyses. *Gigascience*. 2016. 5 :25. Impact Factor : 7.463

— (2019-01) Sempéré G, Pétel A, Rouard M, Frouin J, Hueber Y, De Bellis F, **Larmande P**. Gigwa v2 – Extended and improved genotype investigator. *GigaScience*, Volume 8, Issue 5, May 2019, giz051 Impact Factor : 7.31

— (2019-02) Nti-Addae Y, Matthews D, Ulat V, Syed R, Sempere G, Petel A, Renner J, **Larmande P**, Guignon V, Jones E, Robbins K. Benchmarking Database Systems for Genomic Selection Implementation. 2019. *Database.* pii : baz096. Impact Factor : 3.978

— (2019-03) Abbeloos R, Backlund JE, Basterrechea Salido M, Bauchet G, Benites-Alfaro O, Birkett C, Calaminos VC, Carceller P, Cornut G, Vasques Costa B, Edwards JD, Finkers R, Gao SY, Ghaffar M, Glaser P, Guignon V, Hok P, Kilian A, König P, Lagare JEB, Lange M, Laporte MA, **Larmande P**, LeBauer D, Lyon D, Marshall D, Matthews D, Milne I, Mistry N, Morales N, Mueller L, Neveu P, Papoutsoglou E, Pearce B, Perez-Masias I, Pommier C, Ramirez-Gonzalez RH, Rathore A, Raque AM, Raubach S, Rife T, Robbins K, Rouard M, Sarma C, Scholz U, Selby P, Sempéré G, Shaw P, Simon R, Soldevilla N, Stephen G, Sun Q, Tovar C, Uszynski G, Verouden M. BrAPI - an Application Programming Interface for Plant Breeding Applications. 2019. *BioInformatics.* pii :btz190. Impact Factor : 5.41

4.4 Vers de nouvelles approches d'intégration sémantique des données agronomiques

Contexte et Problématique Entre 2013 et 2017, j'ai été impliqué dans le projet « Institut de Biologie Computationnelle » (IBC) et ai été coordinateur de l'axe « intégration des données et connaissances biologiques » qui reprenait les problématiques d'intégration de données pour la biologie des plantes.

Nous avons, dans un premier temps, contribué au développement de méthodes automatiques d'intégration de bases de données biologiques en associant les logiciels WebSmatch [40], développé par l'équipe INRIA Zénith, Bioportal [158, 140] et Biosemantic [233]. Un prototype en a été réalisé [31]. Ce travail a permis d'identifier un besoin dans l'annotation sémantique des données et la gestion des ontologies pour le domaine des plantes.

Par la suite, nous avons cherché à répondre aux questions scientifiques suivantes :

— Gérer les ontologies du domaine agronomique et leur évolution, en permettant un accès centralisé à ces ressources ;

— Identifier des correspondances entre concepts de différentes ontologies ;
— Annoter sémantiquement les données avec des concepts ontologiques ;
— Gérer efficacement des grands volumes de données biologiques pour en extraire de la connaissance et se reposer pour cela sur les technologies du Web sémantique ;
— Exploiter les méthodes d'ingénierie des connaissances via l'utilisation d'ontologies, pour formuler des hypothèses de recherche permettant de lier le génotype au phénotype ;
— Identifier les gènes clés pour l'amélioration des plantes parmi des centaines de résultats potentiels et déterminer les effets des variations génétiques sur l'expression phénotypique dans des populations étudiées.

Afin de répondre à ces questions, nous avons développé deux plateformes dédiées aux connaissances, la première centrée sur les ontologies du domaine agronomique (AgroPortal) et la deuxième centrée sur les données intégrées sémantiquement à partir de ressources du domaine (AgroLD).

4.4.1 La plateforme AgroPortal

Avec Clément Jonquet (MdC Lirmm), nous avons réalisé un premier prototype d'entrepôt d'ontologies pour le domaine de l'agronomie, nommé AgroPortal. Une première version de la plateforme est d'ores et déjà déployée et maintenue sur un serveur du LIRMM [7].

AgroPortal [98, 96] (2018-04) reprend la technologie du NCBO BioPortal (portail pour la santé et les ontologies biomédicales [8]). Cette technologie est open-source et indépendante du domaine thématique concerné. Le portail propose des services de recherche d'ontologies et de visualisation, avec possibilité de déposer des commentaires et des notes. Le portail offre également un service d'annotation sémantique de données avec les ontologies. L'objectif principal de ce projet est de permettre une utilisation simple des ontologies liées au domaine de l'agronomie, en proposant aux chercheurs de prendre en charge les questions d'ingénierie des connaissances complexes pour annoter les données de recherche. De nombreuses contributions scientifiques ont été réalisées pour améliorer les fonctionnalités du portail. Des nouvelles méthodes de scores [140] ont été développées pour classer les mappings avec des termes ontologiques. Un nouvel algorithme de recommandation a été implémenté dans le *Recommender* [180] et un nouveau modèle de métadonnées a été développé et implémenté dans la plateforme [212].

Mes contributions portent sur la gestion d'un ensemble d'ontologies utilisées pour annoter les données du riz dans le cadre du projet CRP-RICE. Dans ce cadre, je maintiens plusieurs vues (CRP-Rice et AgroLD) qui contiennent un sous-ensemble des ontologies d'AgroPortal.
Récemment, j'ai pu encadrer un stage de master2 (Djibril Kazim) sur le développement d'un pipeline d'annotation sémantique utilisant les fonctionnalités d'AgroPortal pour créer des annotations entre les données d'AgroLD et les ontologies d'AgroPortal. Ces annotations seront utiles pour associer des données issues de sources hétérogènes. Dans un premier temps, nous avons développé un logiciel, Table2Annotation [116] permettant d'identifier des concepts d'ontologies dans les données de tableurs (csv, Excel). Des outils d'annotation ont également été développés dans les pipelines de transformation de données en RDF (voir la section suivante). Je maintiens également, au niveau d'AgroPortal, les liens de correspondances entre les ontologies et les données qui sont importées dans AgroLD.

7. `http://agroportal.lirmm.fr`
8. `http://bioportal.bioontology.org`

4.4.2 La plateforme AgroLD

Contexte AgroLD [9] [220] (2018-1) est une base de connaissance utilisant les technologies du Web sémantique comme structure pour intégrer les données. Elle est conçue pour intégrer des informations disponibles sur diverses espèces végétales du domaine agronomique telles que les espèces de riz (du genre Oryza), Arabidopsis, le blé et le sorgho. Lorsque le projet a démarré, il existait des projets équivalents dans le domaine biomédical et bioinformatique Bio2RDF [16, 30], EBI RDF [100], ou encore Uniprot RDF [174] mais aucun dans le domaine agronomique. Avec l'aide d'un post-doctorant (Dr. Aravind Venkatesan) recruté dans le cadre du projet IBC en 2014, nous avons élaboré des modèles de données et développé un premier prototype.

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 [11, 207], une ontologie sur la fonction des gènes, Plant Ontology [169], une ontologie sur l'anatomie des plantes, Plant Trait Ontology [42], une ontologie sur les caractères phénotypiques des plantes, Plant Environment Ontology [28], une ontologie sur l'environnement en interaction avec les plantes. La majorité de ces ontologies sont hébergées par le projet OBO Foundry [198]. 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 [206] qui identifie les gènes chez les plantes cultivées , UniProtKB [134] qui répertorie les protéines et leurs fonctions chez tous les êtres vivants, Gene Ontology Annotation [13] 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 [10] comme TropGeneDB [82], une base de données de génétique chez les plantes tropicales, OryGenesDB [56], une base de données de génomique chez le riz, GreenPhylDB [181], une base de données de génomique comparative chez les plantes tropicales, OryzaTagLine [114], une base de données de phénotype chez le riz et SniPlay [52], une base de données de variations génomique chez le riz. Ces ressources regroupent des données expérimentales produites par les chercheurs montpelliérains et leurs partenaires. Le tableau 4.1 donne un aperçu des espèces et sources intégrées.

Contributions dans le domaine de l'ingénierie des connaissances

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 [11] ou csv2rdf4lod [12], RML [13], 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) [14] qui représente les données génomique dans un format de type tsv (fichier avec des tabulations comme séparateurs).

9. http://www.agrold.org
10. http://southgreen.fr/
11. https://project.inria.fr/datalift
12. http://purl.org/twc/id/software/csv2rdf4lod
13. https://rml.io/
14. http://gmod.org/wiki/GFF3

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.
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 le GFF, le Gene Ontology Annotation File (GAF) [15], le Variant Call Format (VCF) [48] et travaillons actuellement à packager ces modèles dans une API [16]. 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, notamment pour les données phénotypiques, en fonction des cas d'utilisations que nous réaliserons.

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. Fin 2019, la base de connaissances AgroLD contenait environ 100 millions de triplets RDF créés en convertissant plus de 50 jeux de données provenant de 10 sources de données. 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 GO ou TO (i.e. GO :0005524 est transformé en URI [17]). Dans ce cas, nous avons généré des propriétés supplémentaires avec les ontologies correspondantes, ajoutant ainsi 22% de triplets supplémentaires validés manuellement (voir les détails dans le tableau 4.1). 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.

De plus, nous avons utilisé l'API de service Web AgroPortal pour enrichir les données en annotations sémantiques. Par exemple, pour extraire l'URI correspondant au taxon disponible pour certains standards de données tels que GFF. Mais également pour identifier des concepts ontologiques dans les données comme l'organe d'une plante (e.g. leaf est annoté avec PO_0025034 [18]) ou un caractère phénotypique (plant height serait annoté avec le concept ayant pour URI [19]). Comme le montre la figure 4.4, le workflow de transformation utilise AgroPortal pour annoter les données au moment de la transformation. De plus, nous avons développé une application spécifique pour traiter les formats de fichiers semi-structurés (tsv, csv, excel) [20] et mieux contrôler l'annotation sémantique faite par AgroPortal et y gérer les différentes annotations pour un résultat optimal.

Méthodes de liage d'entités issus de graphes distincts Les graphes RDF partagent un espace de noms commun [21] 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 [112] qui fournit des patrons de conception pour chaque source enregistrée. Par exemple, les gènes intégrés à partir de Ensembl plant sont

15. http://geneontology.org/page/go-annotation-file-format-20
16. https://github.com/pierrelarmande/AgroLD-ETL
17. http://purl.obolibrary.org/obo/GO_0005524
18. http://purl.obolibrary.org/obo/PO_0025034
19. http://purl.obolibrary.org/obo/TO_0000207
20. https://github.com/pierrelarmande/ontology-project
21. http://www.southgreen.fr/agrold/

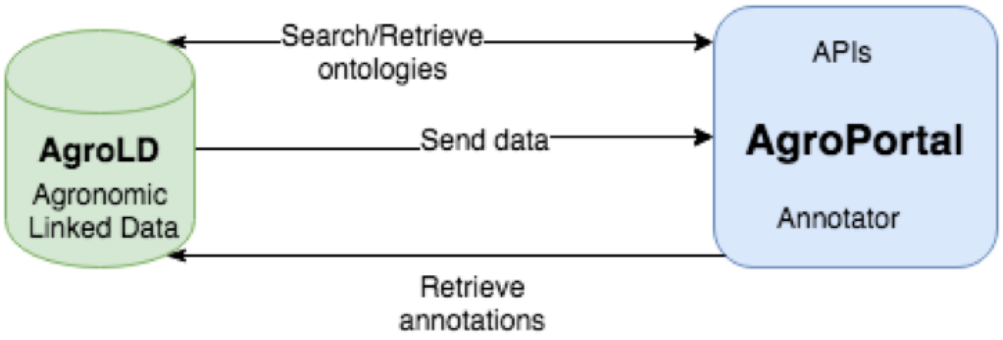


FIGURE 4.4 – Processus d'annotation sémantique entre AgroPortal et AgroLD. Dans une première étape AgroLD utilise le Recommender d'AgroPortal pour identifier les ontologies nécessaires pour annoter le texte. Puis, les éléments du texte sont envoyés à l'API d'Annotator avec des paramètres correspondants dont les ontologies sélectionnées. L'Annotator renvoie les résultats qui sont intégrés dans le pipeline d'annotation.

identifiés par l'URI de base [22]. Lorsqu'elles ne sont pas fournies par Identifiers.org, de nouvelles URI sont construites ; dans ce cas, les URI prennent la forme [23]. Par ailleurs, les propriétés reliant les entités sont construites sous la forme [24].
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 (voir deuxième partie du mémoire). Enfin, nous avons évalué les différents standards de provenance pouvant être utilisés pour annoter les modèles RDF et les annotations sémantiques (2017-02).

Afin de faire correspondre les différents types de données et propriétés, nous avons développé un schéma [25] qui associe les classes et propriétés identifiées dans AgroLD avec des ontologies correspondantes. Par exemple, la classe Protein [26] est associée à la classe polypeptide [27] 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* [28], 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* [29], relie les entités aux termes TO équivalent. Elle est associée à une propriété plus générique. Pour l'instant, 55 mappings ont été identifiés.

22. http://identifiers.org/ensembl.plant/
23. http://www.southgreen.fr/agrold/[resource_namespace]/[identifier]
24. http://www.southgreen.fr/agrold/vocabulary/[property]
25. https://github.com/SouthGreenPlatform/AgroLD
26. http://www.southgreen.fr/agrold/resource/Protein
27. http://purl.obolibrary.org/obo/SO_0000104
28. http://www.southgreen.fr/agrold/vocabulary/has_function
29. http://www.southgreen.fr/agrold/vocabulary/has_trait

Sources de données	Format de fichier	Nb Tuples	Espèces	Ontologies utilisées	Nb de triplets produits
Oryzabase	Custom flat file	17K	R	GO,PO,TO	153K
GO associations	GAF	1, 160K	All	GO	2, 700K
OryGenesDB	GFF	1, 100K	R, S, A,	GO, SO	2, 300K
Gramene	Custom flat file	1, 718K	All	All	5, 172K
UniprotKB	Custom flat file	1, 400K	All	GO, PO	50, 000 K
Oryza Tag Line	Custom flat file	22K	R	PO, TO, CO	300K
TropGeneDB	Custom flat file	2K	R	PO, TO, CO	20K
GreenPhylDB	Custom flat file	100K	R,A	GO, PO	700K
SNiPlay	HapMap, VCF	16K	R	GO	16,000K
Q-TARO	Custom flat file	2K	R	PO, TO	20K
TOTAL					**87,400K**

TABLE 4.1 – **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)

Comment 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) (2017-01). Ainsi la plateforme AgroLD fournit 4 points d'entrée :

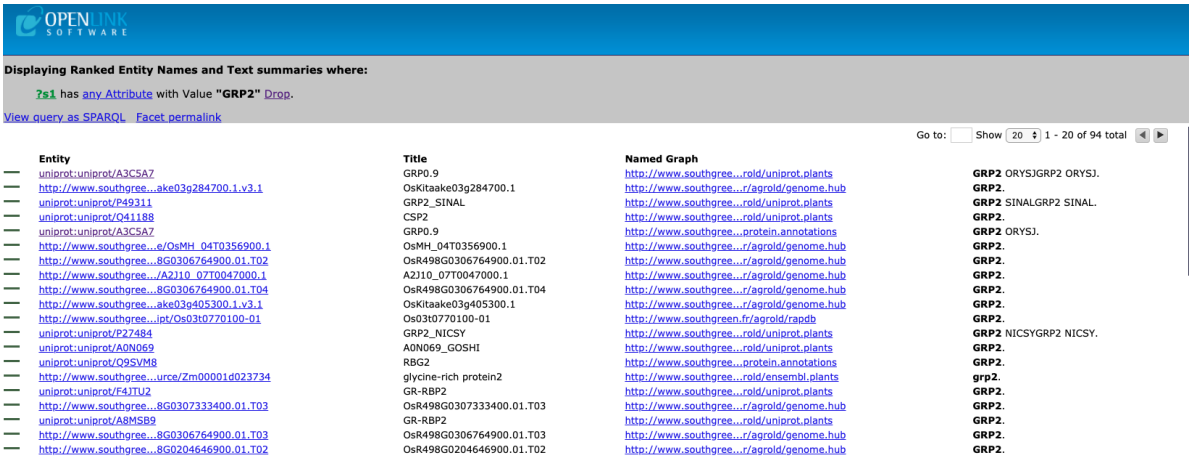


FIGURE 4.5 – Résultat du Quick Search

— **Quick Search** [30], un plugin de recherche à facette mis à disposition par Virtuoso, qui permet

30. http://www.agrold.org/quicksearch.jsp

aux utilisateurs d'effectuer des recherches par mots-clés et de parcourir le contenu d'AgroLD en naviguant dans les liens (Figure 4.5) ;

— **SPARQL Editor**[31], un éditeur de requêtes SPARQL qui fournit un environnement interactif pour la formulation de requêtes SPARQL (voir figure 4.6). Avec un étudiant de Master 2 (Gildas Tagny), nous avons développé l'éditeur en se basant sur les outils YASQE et YASR [175] et l'avons adapté pour notre système.

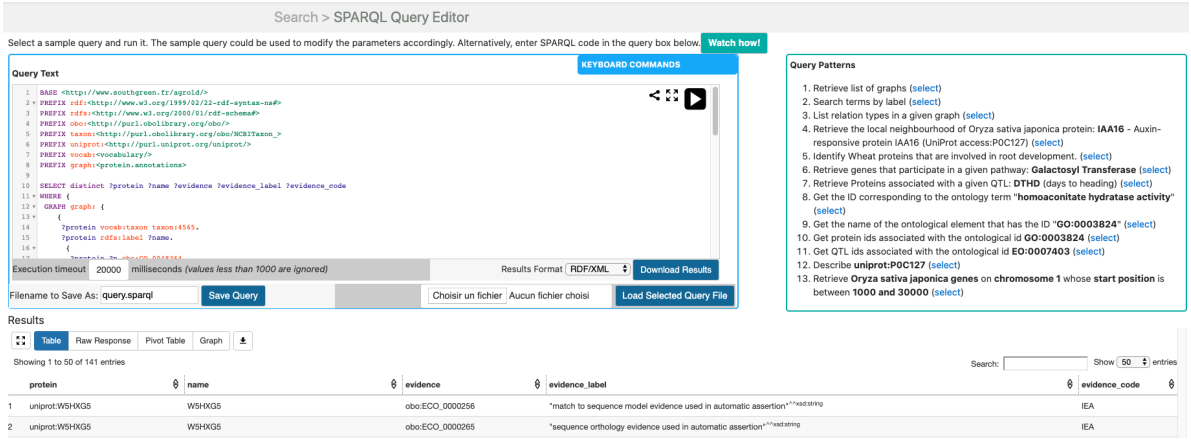


FIGURE 4.6 – Résultat de SPARQLEditor

Le langage SPARQL est un outil puissant pour extraire des informations utiles de la base de connaissances. Par exemple, sur une requête simple : *Identify wheat proteins that are involved in root development (ontology term)*.
Dans le premier exemple de requête Q1 (cf. Figure 4.7) aux lignes 25 et 32, nous utilisons une simple variable $?p$ pour identifier la propriété qui est en relation avec l'objet obo :GO. Dans ce cas le résultat obtenu contient 73 entrées.
Alors que dans le deuxième exemple de requête Q2 (cf. Figure 4.8) aux lignes 23 et 26, nous utilisons un *property path* pour identifier les propriétés qui ont une relation ou une sous-relation avec l'objet obo :GO. Dans ce cas le résultat obtenu contient 137 entrées.

Par ailleurs, le langage SPARQL étant plus expressif, il est possible de composer des requêtes complexes recherchant sur plusieurs graphes. Toutefois, car SPARQL est difficile à appréhender pour les utilisateurs non avertis, 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. Accessoirement, des outils fonctionnels ont été ajoutés comme la possibilité d'enregistrer la requête et de télécharger les résultats dans divers formats tels que JSON, TSV et RDF / XML. De plus, les requêtes créées par l'utilisateur peuvent également être chargées dans l'éditeur.

— **Explore Relationships**[32], est une version modifiée de RelFinder [86] qui permet aux utilisateurs d'explorer et de visualiser les relations existantes entre entités (voir figure 4.9).
Les relations entre les objets constituent une information importante, d'où l'idée d'effectuer des recherches qui permettent à partir d'un point de départ d'explorer le graphe, ou d'éliminer des parties du graphe des données à l'aide de filtres en fonction des besoins de l'utilisateur. L'application RelFinder découvre automatiquement de telles relations dans une source

31. http://www.agrold.org/sparqleditor.jsp
32. http://www.agrold.org/relfinder.jsp

AgroLD

```
### Identify Wheat proteins that are involved in root development. ###

#S3_Q1 - Query without using poperty path

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

SELECT distinct ?protein ?name ?evidence ?evidence_label ?evidence_code
WHERE {
 GRAPH graph: {
    {
     ?protein vocab:taxon taxon:4565.
     ?protein rdfs:label ?name.
      {
       ?protein ?p obo:GO_0048364.
       ?protein vocab:has_annotation ?bp.
       ?bp rdf:subject ?protein.
       ?bp rdf:object obo:GO_0048364.
       ?bp vocab:evidence_code ?evidence_code.
       ?bp vocab:evidence ?evidence.
      } UNION {
       ?protein ?p obo:GO_2000280.
       ?bp rdf:subject ?protein.
       ?protein vocab:has_annotation ?bp.
       ?bp rdf:object obo:GO_2000280.
       ?bp vocab:evidence_code ?evidence_code.
       ?bp vocab:evidence ?evidence.
      }
    }
 }
    GRAPH ?g {
    ?evidence rdfs:label ?evidence_label.

    }
}

# RESULTS = 73 entries
```

FIGURE 4.7 – Exemple de requête SPARQL - Q1

de données disposant d'un serveur d'accès SPARQL et les affiche sous forme de graphe. Elle aide l'utilisateur à trouver les classes d'entités avec une fonctionnalité d'auto-complétion. Cependant, la version d'origine de RelFinder a été développée (en ActionScript) et configurée pour DBpedia. Nous avons proposé une configuration et une modification du système adapté à AgroLD. La configuration concerne principalement le point d'accès SPARQL, les

AgroLD

```
### Identify Wheat proteins that are involved in root development. ###

#S3_Q2 - Query using Property path


BASE <http://www.southgreen.fr/agrold/>
PREFIX rdf:<http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX rdfs:<http://www.w3.org/2000/01/rdf-schema#>
PREFIX obo:<http://purl.obolibrary.org/obo/>
PREFIX taxon:<http://purl.obolibrary.org/obo/NCBITaxon_>
PREFIX uniprot:<http://purl.uniprot.org/uniprot/>
PREFIX vocab:<vocabulary/>
PREFIX graph2:<protein.annotations>
PREFIX graph1:<go>
SELECT distinct ?protein ?name ?label ?evidence ?evidence_label ?evidence_code
WHERE {
  GRAPH graph1: {
    {
      ?term rdfs:subClassOf* obo:GO_0048364.
      ?term rdfs:label ?label.
  } UNION {
      ?term rdfs:subClassOf* obo:GO_2000280.
    ?term rdfs:label ?label.
      }
    }
  GRAPH graph2: {
      ?protein vocab:taxon taxon:4565.
      ?protein rdfs:label ?name.
      ?protein ?p ?term.
      ?protein vocab:has_annotation ?bp.
      ?bp rdf:subject ?protein.
      ?bp rdf:object ?term.
      ?bp vocab:evidence_code ?evidence_code.
      ?bp vocab:evidence ?evidence.
    }
  GRAPH ?g {
    ?evidence rdfs:label ?evidence_label.

    }
}

# RESULT = 137 entries


```

FIGURE 4.8 – Exemple de requête SPARQL - Q2 avec Property Path

propriétés à prendre en compte pour la recherche d'entités et la description des ressources. De plus, nous avons ajouté quelques exemples biologiques pour guider les utilisateurs ;

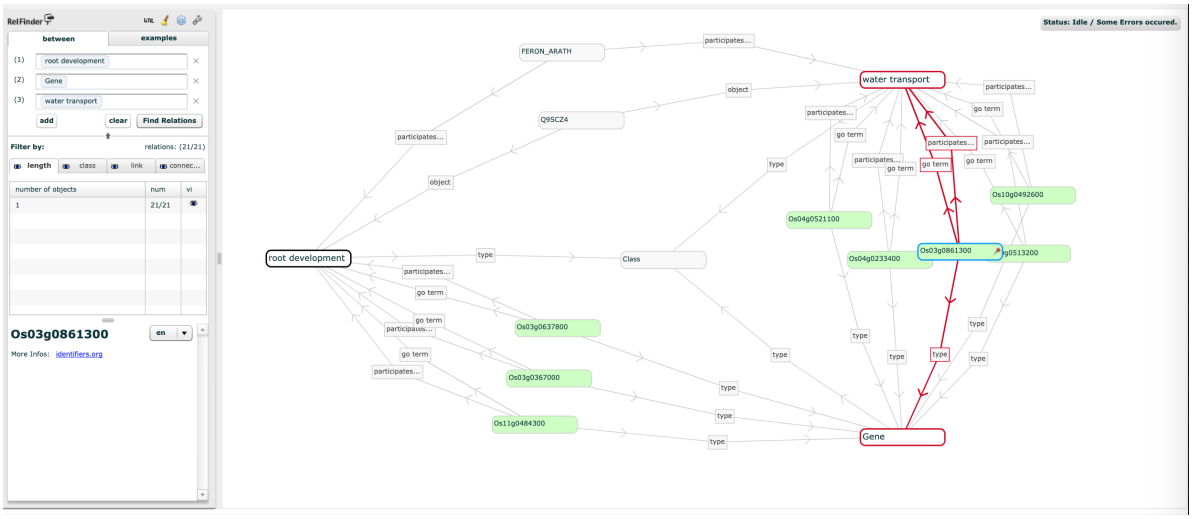


FIGURE 4.9 – Résultat du Explore Relationships

— **Advanced Search** [33], un formulaire proposant des recherches spécifiques par entité et possédant un moteur d'agrégation de ressources externes (voir figure 4.10).

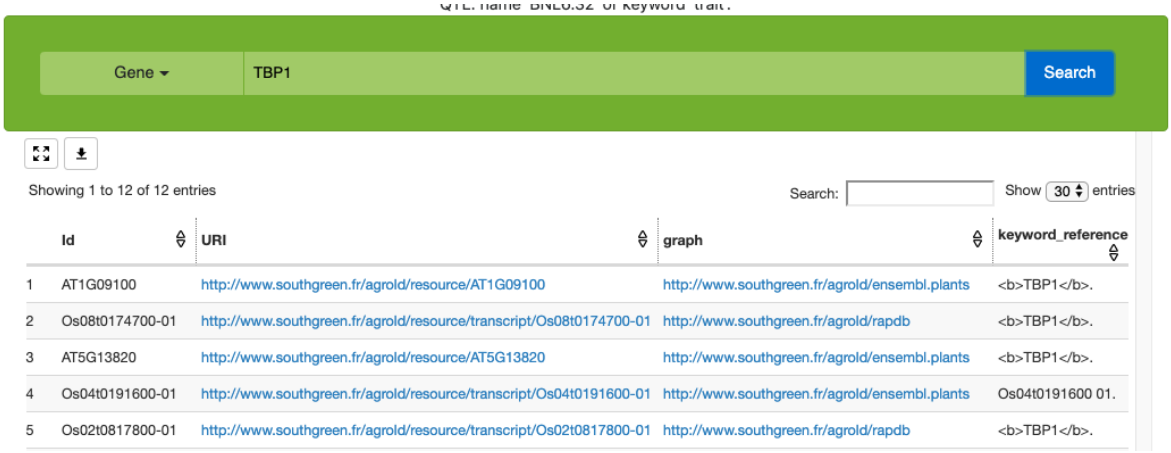


FIGURE 4.10 – Résultat de l'Advanced Search

Le formulaire Advanced Search est basé sur une API REST [34], entièrement développée dans le cadre du stage de master 2 AgroLD, par Gildas Tagny. Le but de ce formulaire est de fournir aux biologistes un outil permettant d'interroger la base de connaissances tout en masquant les 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.

Le projet AgroLD nous a permis d'identifier de nombreux challenges sur le plan informatique ouvrant de nouvelles pistes de travail. Cette première phase de travail a été publié récemment (2018-01a)

33. http://www.agrold.org/advancedSearch.jsp
34. http://www.agrold.org/api-doc.jsp

Exploration de méthodes de ré-écriture du langage naturel vers SPARQL Toujours dans l'objectif de réduire la barrière de langage pour interroger la base de connaissances, nous avons évalué des systèmes de question-réponses (SQR) pour la traduction de la langue naturelle en SPARQL. Ce sont des systèmes permettant aux utilisateurs de poser des questions en langage naturel et de leur donner des réponses concises [88, 131]. Actuellement, les systèmes SQR sont utilisés dans divers domaines et peuvent également être une solution prometteuse pour la biologie végétale. Dans le domaine médical, plusieurs travaux ont été menés qu'il était intéressant d'évaluer, d'exploiter, voire d'étendre. Nous avons développé un test de référence (*Gold Standard*[35]) afin d'évaluer ces systèmes car les données agronomiques étaient absentes des *gold standard* actuels (2016-05). Nous avons regroupé différentes informations de la littérature [88, 131, 147, 153] pour mettre en place une classification des systèmes SQR en fonction des principales approches explicitées précédemment (illustrée à la figure 4.11 et effectué une évaluation empirique de ces systèmes. Finalement,

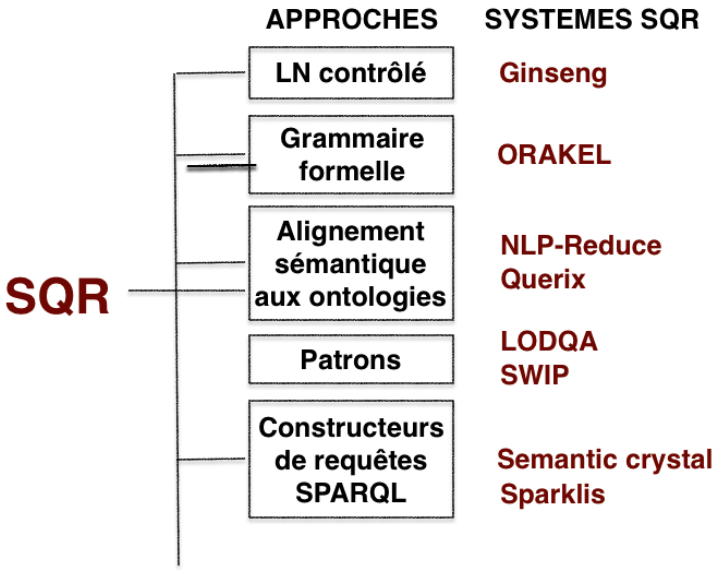


FIGURE 4.11 – Vue générale de notre classification SQR

nous avons porté notre attention sur le système LODQA [105]. Le système est basé sur 3 modules i) un module de décomposition de la phrase en langage naturel, ii) un module de correspondance de termes, iii) un module de réécriture en requête SPARQL. LODQA étant très dépendant du domaine, c'est-à-dire que les mots identifiables sont indexés dans le module 2, les tests que nous avons réalisés n'ont pas donné de bons résultats. Il n'a pas été possible durant le stage de contribuer sur le module de correspondance de termes.

4.4.3 Conclusion

En conclusion de cette section, nous pouvons dire que l'utilisation des technologies du Web sémantique semble être adaptée et adoptée par une large communauté du domaine des sciences de la vie. A travers nos projets AgroPortal et AgroLD, nous apportons notre contribution pour le domaine agronomique. Pour cela, nous participons aux développement de méthodes de transformation RDF de standards de données et fournissons des outils réutilisables. Nous proposons également des modèles de représentation des connaissances à travers des schémas RDF ou OWL.

35. *Gold Standard* : indique le meilleur test du moment permettant d'évaluer une méthode, dans notre cas il s'agissait d'évaluer les méthodes sur des données agronomiques qui étaient absentes des *gold standard* actuels.

En ce qui concerne l'accès aux données liées, nous contribuons aux développements de méthodes et interfaces facilitant leur utilisation par un plus grand nombre d'utilisateurs notamment à destination des biologistes. Toutefois, il reste encore à répondre à de nombreux challenges. En effet, pour permettre une meilleure intégration des graphes entre-eux, de nouvelles méthodes de liage de données doivent être développées. Nous constatons aussi que l'extraction de connaissances dans les données non-structurées pourrait grandement améliorer le processus d'intégration. Finalement, l'utilisation de la sémantique des données pour inférer de nouvelles connaissances par l'intermédiaire du raisonnement sur les données et schéma devrait être mise en œuvre et généralisée.

Sélection de références

— (2020-1) **Larmande P**., Jibril K. Enabling Fast Annotation Process With Table2Annotation Tool. *bioRxiv* 2020.04.03.023069 ; doi : https ://doi.org/10.1101/2020.04.03.023069

— (2018-01) Venkatesan A., Tagny G., El Hassouni N., Chentli I., Guignon V., Jonquet C., Ruiz M., and **Larmande P**. Agronomic Linked Data (AgroLD) : a Knowledge-based System to Enable Integrative Biology in Agronomy. *PLoS ONE* 13(11) : e0198270. Impact Factor : 2.766

— (2018-02) Do H., Than K., and **Larmande P**. Evaluating Named-Entity Recognition approaches in plant molecular biology. MIWAI 2018. *Sringer LNAI proceedings* 11248. pp 219-225. 2018

— (2018-03) **Larmande P.**, El Hassouni N. , Venkatesan A., Tagny G., Ruiz M. The Agronomic Linked Data project (AgroLD) : a knowledge network platform for rice. *Oral presentation at International Symposium on Rice Functional Genomics ISRFG* 2017. Sewon (Korea)

— (2018-04) Jonquet C., Toulet A. , Arnaud E., Aubin S. Dzalé Yeumo E., Emonet V., Graybeal J., Laporte M. A., Musen M. A. Pesce V. and **Larmande P**. AgroPortal : A vocabulary and ontology repository for agronomy. *Computers and Electronics in Agriculture*. 2018 ;144 ;126-143 Impact Factor : 2.201

— (2017-07) **Larmande P**., El Hassouni N. , Venkatesan A., Tagny G., Ruiz M. The Agronomic Linked Data project (AgroLD) : a knowledge network platform for rice. *Oral presentation at International Symposium on Rice Functional Genomics ISRFG* 2017. Sewon (Korea)

— (2017-06) Venkatesan A., Tagny G., El Hassouni N., Ruiz M., **Larmande P**. The Agronomic Linked Data project. *Computer demo at Plant and Animal Genomes Conference PAG* 2017. San Diego, (USA).

— (2017-04) Dzale Yeumo E, Alaux M, Arnaud E, Aubin S, Baumann U, Buche P, .. **Larmande P** et al. Developing data interoperability using standards : A wheat community use case. *F1000Research*. 2017 ;6 :1843.

— (2017-02) Cohen-Boulakia S, Belhajjame K, Collin O, Chopard J, Froidevaux C, Gaignard A, .. **Larmande P** et al. Scientific workflows for computational reproducibility in the life sciences : Status, challenges and opportunities. *Futur. Gener. Comput. Syst.* 2017 ;75. Impact Factor : 2.786

— (2017-01) Ngompé GT, Venkatesan A, Hassouni N, Ruiz M, **Larmande P**. AgroLD API Une architecture orientée services pour l'extraction de connaissances dans la base de données liées AgroLD. *Ingénierie des Systèmes d'Informations Ed. Lavoisier.* 2016. 21 :133–58. Impact Factor : 1.046

— (2016-03) Jonquet C, Toulet A, Arnaud E, Aubin S, Yeumo ED, Emonet V, Graybeal J, Musen MA, Pommier C, **Larmande P**. 2016. D202 : Reusing the NCBO BioPortal technology for agronomy to build AgroPortal. *Oral Presentation at International Conference on Biomedical Ontology and BioCreative ICBO BioCreative* 2016. Corvalis (USA)

— (2016-04) Zevio S., El Hassouni N., Ruiz M. and **Larmande P**. AgroLD indexing tools with ontological annotations. *Poster at Semantic Web for Life Science SWAT4LS* 2016. Cambridge (UK)

— (2016-05) Chentli I, **Larmande P** et Todorov K. Construction d'un gold standard pour les données agronomiques. *IC2016*, Montpellier, France.

— (2016-06) Robakowska Hyzorek D, Mirouze M, **Larmande P**. Integration and Visualization of Epigenome and Mobilome Data in Crops. *Journées ouvertes pour la Biologie, l'informatique et les Mathématiques (JOBIM)*. Lyon, 2016.

Les articles illustratifs correspondant aux activités et résultats de recherche seront présentés en annexe.

Deuxième partie

Projet et perspectives

Chapitre 5

Projet

L'intégration des données et l'extraction de connaissances en biologie, comme nous l'avons relaté dans le chapitre 4 ont toujours été au cœur des travaux effectués afin, notamment dans le cas du riz pris comme modèle, d'identifier les gènes clés pour l'amélioration des plantes.
Ce chapitre présente les perspectives ouvertes par les activité déjà menées. Seront détaillées les directions et évolutions des recherches pressenties dans les prochaines années et bien sur corrélées aux étudiants que je supervise et aux projets auxquels je participe. Elles constituent le projet de recherche dont l'objectif est d'approfondir la représentation des informations diverses dont nous disposons sous forme de **graphes de connaissances** afin de formuler de nouvelles hypothèses de recherche permettant de lier le génotype au phénotype.
Seront présenté dans les diverses sections :

— l'axe dédié à l'intégration de données et l'extension de connaissances précisera les évolutions envisagées pour renforcer l'intégration de données hétérogènes dynamiquement dans un graphe de connaissance tout en permettant un élargissement des connaissances associées par annotation sémantique ;

— l'axe dédié à l'extraction et à l'enrichissement des connaissances qui précisera l'obtention de graphes de connaissances à partir de méthodes d'extraction, de liage et de raisonnements ;

— l'axe dédié à l'exploitation des graphes de connaissance pour rechercher, identifier les éléments porteur d'amélioration pour les plantes ; il précisera les hypothèses relatives aux méthodes de priorisation des gènes aidant à l'analyse fonctionnelle des réseaux d'interactions moléculaires.

5.1 Intégration de données et annotation sémantique

5.1.1 Intégration dynamique des données

La première étape du développement du graphe de connaissance sera d'intégrer et de transformer en RDF de nouvelles ressources pour le riz afin de construire un large réseau d'interactions moléculaires (réseaux de co-expression de gènes, transcriptomique, facteurs de transcriptions, complexes proteine-proteine, etc.). Ce processus de transformation est souvent appelé le **lifting de données** [1].

Je m'appuierai sur le projet AgroLD, une base de connaissance que je développe activement depuis 2015. Dans ce contexte, j'ai eu l'occasion de développer de nombreux outils de *lifting* soit pour des sources de données spécifiques, développées au sein de la communauté (voir section 4.4.2), soit pour des formats de fichier génériques (voir section 4.4.2). Dans ce cas, l'intérêt d'intégrer de nombreuses ressources est de pouvoir agréger des données complémentaires dans un même modèle de représentation, ici RDF, afin de réduire l'hétérogénéité syntaxique.

1. Le processus de « lifting » des données (conversion, publication et interconnexion) s'appuie sur des vocabulaires contrôlés possédant une sémantique formelle, en d'autres termes des ontologies

Une attention particulière sera portée aux données expérimentales produites par les chercheurs de l'unité Diade et les partenaires des pays du sud. Travailler sur des données expérimentales présente divers intérêts scientifiques :

— D'un point de vue informatique, il y a un réel enjeu de confronter les méthodes et algorithmes issus de l'état de l'art, souvent développés à partir de données exemples (petit jeux de données présentant une faible représentativité de la réalité), sur ce type de données de terrain qui sont hétérogènes, bruitées, incomplètes et complexes. De plus s'agissant souvent de données volumineuses (par exemple les données de génotypages ou des images), la méthode d'intégration doit éviter de transformer intégralement les données afin de garantir de bonnes performances.

— D'un point de vue biologique, il y a un fort intérêt d'intégrer les données expérimentales et de les combiner aux données « validées », extraites de publications ou de bases de données, afin d'enrichir la connaissance sur les résultats obtenus.

Pour l'extraction d'information contenue dans des bases de données relationnelles, nous avons développé une application BioSemantic, au cours de la thèse de Julien Wollbrett. Cette dernière propose une approche flexible et automatisée pour la création de vues RDF sur des bases de données et assiste l'utilisateur dans la formulation de requêtes se basant sur ces vues en développant un algorithme de recherche du plus court chemin dans les graphes RDF [233] (plus de détail sur cette méthode sont donnés en section 4.2). Toutefois, il existe de nombreux outils de transformation et langages de *mapping* associés, adaptés aux différents types de SGBD et aux modèles de représentation des données. L'article de Michel *et al* [142] en dresse un inventaire et compare les différentes méthodes. De plus, les auteurs proposent également l'outil xR2RML [143] qui présente l'avantage de transformer les données à la demande au cours d'une requête et ce pour différents types de bases de données (XML, objet, Relationnel et NoSQL). En collaboration avec l'équipe Wimmics, je compte développer ces aspects pour extraire les données expérimentales de variations génomiques actuellement stockées dans Gigwa [187].

5.1.2 Annotation sémantique

Ainsi, dans cette première phase de transformation, chaque graphe RDF produit est indépendant des autres. C'est grâce aux ontologies que les liens sémantiques entre les entités biologiques peuvent être créés. Dans notre domaine, le cadre conceptuel pour la gestion des connaissances est basé sur des ontologies bien établies : Gene Ontology (GO), Sequence Ontology (SO), Plant Ontology (PO), Trait Ontology (TO), Phenotype quality ontology (PATO) et Environnement Ontology (EnvO). Un lien sémantique (e.g. annotation sémantique) est créé dès lors qu'une entité biologique référence un terme ontologique (e.g. la protéine IAA16 est exprimée dans « le coléoptile » qui a pour URI OBO :PO_0020033). Ainsi, il est possible de relier des entités d'un même graphe ou dans des graphes différents dès lors qu'elles partagent les mêmes liens sémantiques. Dans AgroLD, nous exploitons les annotation sémantiques lorsqu'elles sont explicitement présentes dans les jeux de données (e.g. un gène est annoté dans une ressource avec le terme GO :xxxxx). Dans le domaine bioinformatique cette étape est souvent appelée enrichissement [24, 193, 6]. Ces annotations sont souvent produites à partir de logiciels bioinformatiques souvent basés sur des recherches de similarité basées sur les séquences nucléotidiques. L'article de Blake *et al*, 2013 [23] donne un aperçu des méthodes d'annotation. Cette méthode nous permet de produire 22 % d'annotations supplémentaires. Toutefois, de nouvelles méthodes doivent être développées pour les nombreuses ressources qui ne possèdent pas ces informations ou ne sont pas basées sur des séquences.

Identifier des liens sémantiques dans les données est un élément important pour la construction des réseaux de connaissances dans AgroLD. C'est également un champ disciplinaire très actif dans la communauté informatique [62, 163].

De fait, de nombreuses méthodes sont proposées afin de relier des « termes » [2] issus de textes divers à des labels de concepts [3] issus de différentes ontologies afin d'augmenter la connaissance. Toutefois peu de travaux proposent des méthodes efficaces dans le cas de caractères phénotypiques complexes comme les maladies ou les phénotypes [83]. Voir les exemples présentés ci-dessous.

Dans notre cas, il y a certaines spécificités dont il faut tenir compte :

— Un terme en langue naturelle désignant une entité biologique peut être représenté par son symbole ou son acronyme : par exemple le gène MOC1, la protéine APO1 ;

— Un terme en langue naturelle désignant une entité biologique peut être polysémique et ambigu, donc difficile à annoter ;

— Un terme ou mot correspondant à un phénotype peut faire référence de manière implicite à plusieurs labels de concepts issus de différentes ontologies. Par exemple, le phénotype *Dwarfism* peut être annoté avec le concept *dwarf-like* de l'ontologie PATO (Phenotype And Trait Ontology), mais il va également de pair avec le concept *Tillering* de l'ontologie PO (Plant Ontology) et le concept *Tiller angle* de l'ontologie TO (Trait Ontology) ;

— Un terme ou mot composé, correspondant à un phénotype, peut être annoté à partir de deux ontologies. Par exemple, le phénotype *wrinkled seed* est composé du label du concept *wrinkled* de l'ontologie PATO et du label du concept *seed* de l'ontologie PO.

Pour répondre à ces défis et à d'autres liés à l'analyse de textes biologiques non structurés, les outils d'annotation sémantiques performants reposent souvent sur une utilisation combinée de traitements de texte, de bases de connaissances, de mesures de similarité sémantique et de techniques d'apprentissage automatique [99]. Agroportal [96] vise à développer un portail d'ontologies de référence pour le domaine agronomique. Le portail ambitionne également de proposer plusieurs outils de recherche et d'annotation sémantique. Comme indiqué dans (Jonquet et al, 2018) [96], nous comptons développer un workflow d'annotation entre AgroPortal et AgroLD basé sur des mesures de similarités, le traitement de texte (voir section 5.2.1) et utilisant les fonctionnalités d'AgroPortal pour réaliser l'association des données avec les concepts ontologiques.

5.2 Extraction et exploitation de la connaissance

5.2.1 Extraction d'entités biologiques et de relations

Un des enjeux du projet sera d'enrichir AgroLD à partir des données non-structurées qui sont contenues dans les publications scientifiques et dans des champs textes des bases de données (par exemple les champs « commentaires », « descriptions »). Nombre de ces champs contiennent, des mécanismes moléculaires et génétiques d'intérêts qui sont souvent décrits par des expressions complexes associant des entités biologiques reliées par des relations sémantiques spécialisées (e.g. *Ehd1 and Hd3a can also be down-regulated by the photoperiodic flowering genes Ghd7 and Hd1*).

2. « terme », d'après CNTRL (Centre National de Ressources Textuelles et Lexicales) : Mot ou ensemble de mots ayant, dans une langue donnée, une signification précise et exprimant une idée définie.

3. Un label correspondra à une forme textuelle du concept. Nous renvoyons le lecteur vers les définitions du W3C `https://www.w3.org/standards/semanticweb/ontology`

Dans le domaine de l'extraction de connaissances, une tâche importante consiste à identifier ce que l'on dénomme par entités nommées *named entity* (entités biologiques classées par type), ici par exemple, les entités *Ehd1, Hd3a, Ghd7 et Hd1*), par des méthodes dites Named Entity Recognition (NER). Pour cela, je souhaite évaluer des approches existantes de « text mining » (Natural Language Processing) développées pour le domaine biologique.

Dans un premier temps, nous avons évalué des approches d'extraction de NER, notamment avec les méthodes récentes qui utilisent les méthodes d'apprentissage profond. Dans la suite de la section, nous comparons une approche développée par Habibi *et al* [80] utilisant un modèle LSTM-CRF (Long Short Term Memory model combiné avec Conditional Random Fields) à une approche développée par Basadella *et al* [15] utilisant un dictionnaire d'entités associé à des classificateurs d'apprentissage automatique. Le but de ce travail sera d'identifier la meilleure approche afin de l'intégrer dans un pipeline d'extraction d'information et formalisation de connaissances que les experts puissent valider par l'application AgroLD. Des premiers résultats d'évaluation ont déjà été obtenus, mais ils sont incomplets et nécessitent d'être approfondis. Par la suite, nous comptons également les enrichir en incluant de nouvelles approches issues de l'état de l'art (voir la section 3.5.2).

Afin de pouvoir comparer ces différentes approches, nous avons constitué un corpus de données sur le riz qui pourra servir de modèle d'entraînement pour détecter des entités et leurs relations dans le texte. Ce corpus, *OryzaGP* [115], est composé de plus de 10,000 titres et résumés d'articles scientifiques publiés sur le riz et téléchargés à partir de PubMed. Pour le créer, nous avons utilisé les jeux de données de Oryzabase [4], une base de données intégrée sur le riz. Nous avons téléchargé les jeux de données *Gene List* et *Reference* contenant respectivement une liste de 21 739 gènes connus différents et un ensemble de résumé d'articles avec des gènes associés. Nous avons également fait une annotation manuelle du corpus. Nous avons utilisé le corpus comme données d'entraînement et de validation pour les deux approches d'extraction d'entités.

Évaluation d'approches de NER Pour l'approche proposée par Habibi *et al* [80], l'architecture est basée sur des modèles LSTM-CRF, illustrés à la figure 5.1 [80]. L'ensemble du système comprend trois couches principales : la couche d'intégration en entrée, la couche bi-directionnelle LSTM et la couche CRF en sortie. A partir d'une phrase composée de la séquence de mots w_1 ; w_2 ; ... ; w_n en entrée, la couche d'intégration produit un vecteur d'inclusion x_1 ; x_2 ; ... ; x_n pour chaque mot. Chaque vecteur d'inclusion (ou de plongement, encore nommé *embeddings*) concernant un mot distinct est une concaténation de deux composants : l'inclusion au niveau du mot et du caractère. Nous cherchons les vecteurs d'inclusion de mots à partir d'une table de recherche de vecteurs d'inclusion de mots. En même temps, nous appliquons un LSTM bidirectionnel à la séquence d'inclusion de caractères pour chaque mot, puis concaténons les deux sens pour obtenir l'inclusion au niveau du caractère. Cela signifie que la séquence d'inclusion résultante x_1 ; x_2 ; ... ; x_n est introduite dans la couche LSTM bidirectionnelle afin de produire une représentation plus précise de la séquence d'entrée et passe ensuite en entrée de la dernière couche CRF. La sortie finale de cette couche est obtenue en appliquant l'algorithme de Viterbi [5].

Pour l'approche proposée par Basadella *et al* [15], elle repose sur un dictionnaire d'entités associé à des classificateurs d'apprentissage automatique. Dans un premier temps le dictionnaire de recherche d'entités nommé OGER (OntoGene Entity Recognizer) est utilisé pour annoter les objets dans les ontologies de domaine sélectionnées. Il s'agit d'un service Web qui fournit un accès aux

4. https://shigen.nig.ac.jp/rice/oryzabase/
5. https://en.wikipedia.org/wiki/Viterbi_algorithm

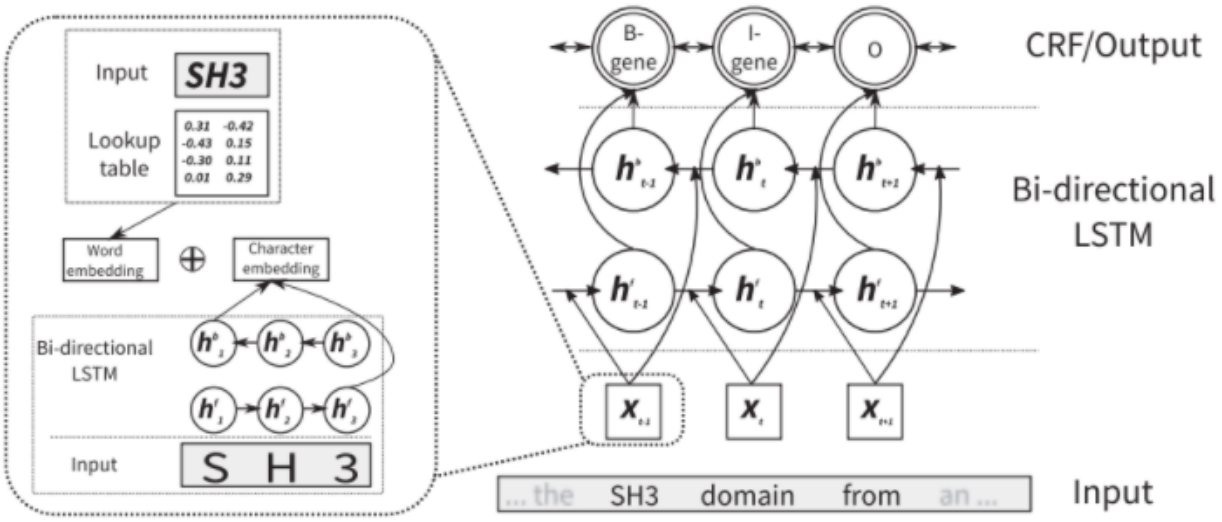


FIGURE 5.1 – Le modèle LSTM-CRF présenté dans [80]

dictionnaires construits sur les bases PubMed du NCBI [6]. Ensuite, le framework *Distiller* est utilisé pour extraire ces informations et leur attribuer des fonctions/types permettant à un algorithme d'apprentissage automatique de sélectionner des entités pertinentes.

Le processus de *Distiller* est basé sur une extraction automatique de mots clés (AKE - *Automatic Keywords Extraction*) pour extraire des informations d'un texte. AKE semble être différent de NER : l'algorithme s'intéresse à la recherche de petits ensembles d'information les plus pertinents dans un document, puis par la recherche de toutes les informations des types sélectionnés. En outre, AKE peut être exécuté à la fois en tant qu'algorithme non supervisé et supervisé, et *Distiller* tire réellement son origine d'une approche non supervisée.
En ce qui concerne son architecture, *Distiller* est organisé en une série de modules, chaque module étant conçu pour effectuer efficacement une tâche unique [14], telle que l'analyse grammaticale de la phrase (part-of-speech tagging), l'analyse statistique, etc. Il fonctionne avec la possibilité d'implémenter différents pipelines pour différentes tâches. Les modules partagent leur informations sur les entités dans une mémoire partagée afin que les autres modules puissent y accéder. L'implémentation d'une tâche d'extraction avec *Distiller* conduit à la spécification d'un pipeline associant les modules. Une tâche d'extraction est normalement divisée en étapes : pré-traitement, sélection de phrase clé candidate et classement de candidats. Le schéma du pipeline *Distiller* est décrit à la figure 5.2.

Pour l'évaluation de cette approche, nous avons mis en œuvre deux algorithmes d'apprentissage automatique différents : les réseaux de neurones (NN) et les CRF. Les performances du *Distiller* sont différentes en fonction du modèle utilisé. Dans le cas de CRF, il utilise la sortie annotée d'OGER en tant que propriété et considère tout élément dans le texte comme une entité à prédire. En revanche, Distiller utilisé seul se concentre uniquement sur le filtrage de la sortie d'OGER et sur le processus de classification pour chaque entité. Pour résumer, nous avons implémenté 3 méthodes pour évaluer l'approche hybride de Basadela *et al* :

— basée sur les résultats OGER ;

6. https ://www.ncbi.nlm.nih.gov/pubmed

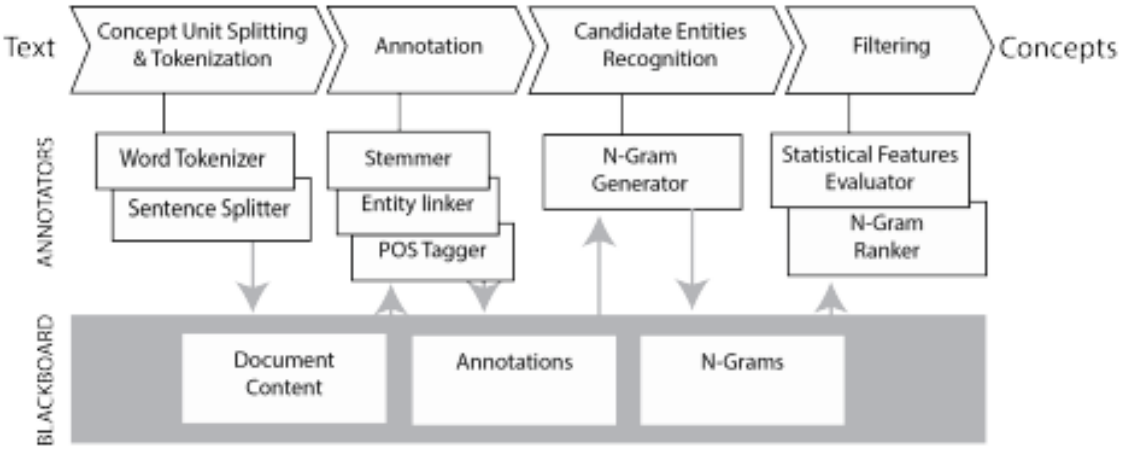


FIGURE 5.2 – Le schéma du Distiller présenté dans [15]

— basée sur un post-traitement des résultats d'OGER avec des réseaux de neurones (NN) ;
— basée sur un post-traitement des résultats d'OGER avec CRF.

Nous avons évalué les performances de tous les modèles sur le jeu de données OryzaGP. Les résultats en termes de précision, rappel, $F_1 - score$ pour chaque modèle sont présentés dans le tableau 5.1.

Nous constatons que la méthode LSTM-CRF obtient de meilleurs résultats. En moyenne, le score F_1 - est de 86,72 % pour la méthode LSTM-CRF et 80,44 % pour la méthode LSTM.

	Precision(%)		Recall(%)		$F_1 - score$(%)	
	(i)	(ii)	(i)	(ii)	(i)	(ii)
LSTM	80.16	78.06	79.16	82.97	79.66	80.44
LSTM-CRF	87.24	87.32	84.73	86.13	85.97	86.72

TABLE 5.1 – Résultats des performances des approches LSTM - le résultat des performances en termes de précision, de rappel et de $F_1 - score$ pour les méthodes LSTM et LSTM-CRF avec différents paramètres d'entraînement : (i) learning rate = 0,001, dropout = 0,3, (ii) learning rate = 0,001, droupout = 0,5

Sur les trois méthodes évaluées avec OGER, le résultat de la méthode OGER est très exploitable (elle se situe autour de 58,5 %) mais c'est la méthode OGER-CRF qui a donné le meilleur résultat, soit 86,72% en moyenne. Le deuxième meilleur résultat a été obtenu avec le modèle OGER associé à un réseau de neurones. Le résultat est présenté dans le tableau 5.2.

Des résultats obtenus à partir des deux approches que nous avons évaluées sur le corpus OryzaGP, nous pouvons dire que l'approche LSTM-CRF obtient un meilleur F1-score que OGER-CRF. Ce sont des résultats encourageants qui nous confortent dans l'utilisation de méthodes d'apprentissage profond pour l'extraction d'entités nommées. A l'avenir, nous comptons élargir l'extraction

	Precision(%)	Recall(%)	$F_1 - score(\%)$
OGER	53.03	65.23	58.50
OG+NN	63.93	71.10	67.32
OG+CRF	88.39	82.24	85.08

TABLE 5.2 – Résultats des performances des approches OGER - le résultat des performances de la méthode hybride en termes de précision, de rappel et de $F_1 - score$

à d'autres types d'entités (dans cet exemple nous n'avions évalué que les entités gènes/protéines) comme les composés chimiques, les noms d'espèces et les caractères phénotypiques. Nous évaluerons également de nouvelles méthodes basées sur l'apprentissage profond (voir section 3.5.2).

5.2.2 Liage des données

Un second élément important dans l'enrichissement de connaissance est le liage (l'interconnexion) de données. Le processus qui peut avoir plusieurs désignations anglophones, *instance matching*, *data linking* et *link discovery* vise à établir des liens sémantiques d'équivalence entre les entités de graphes différents. Il vise à déterminer si deux ressources données se réfèrent ou non au même objet du monde réel. La problématique de liage est un domaine de recherche actif qui a introduit une pléthore d'approches. De fait, de nombreux outils ont été développés pour traiter ce problème au cours des dernières années. Des approches et outils ont été étudiés dans [64, 3, 106].

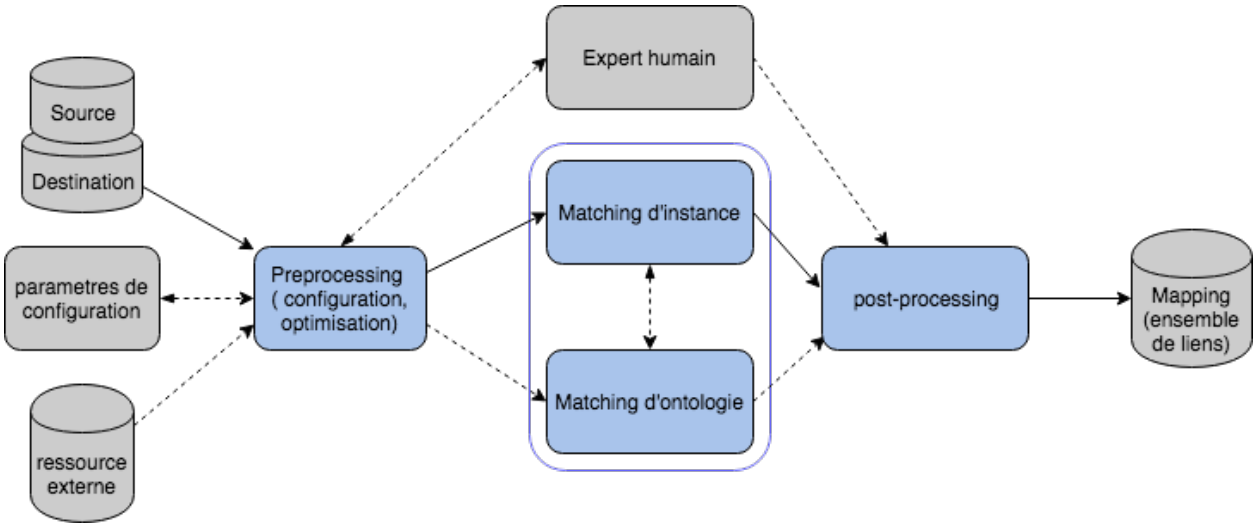


FIGURE 5.3 – Workflow général du processus de liage de données adapté de Achichi et al, 2018 [106]

La majorité des infrastructures de liage actuelles utilisent généralement des workflows composés de plusieurs étapes. Dans la plupart des cas, ces workflows sont des instanciations du workflow générique présenté à la figure 5.3. Les paramètres en entrée incluent en général les deux jeux de données à lier (source, cible), les paramètres de configuration et les ressources externes qui peuvent être facultatives. Les données d'entrée peuvent être fournies sous la forme d'archives RDF / OWL ou sous la forme d'un SPARQL endpoint pour un accès aux données basé sur une requête. Le liage peut être restreint à un sous-ensemble d'une source de données, par exemple des instances d'une classe particulière et il n'est pas nécessaire de les comparer à des entrepôts de données plus génériques tels que DBpedia. Les paramètres de configuration peuvent être des

règles de liage ou des mesures de similarité pour établir des liens d'identité. Les données d'entraînements peuvent être fournies pour des étapes de liage basées sur l'apprentissage. D'autres outils peuvent éventuellement être utilisés comme d'autres sources de connaissance, par exemple, des dictionnaires de données ou des correspondances préalablement définies. La sortie du pipeline correspond à un ensemble des liens trouvés ou des correspondances représentant un liage entre les jeux de données source et cible, en général, déclaré avec des prédicats *owl :sameAs*.

Les challenges du liage de données Il existe de nombreux outils qui tendent à solutionner ce problème mais dans la réalité, le liage de données est un processus complexe et souvent dépendant d'un domaine de connaissance. Dans ce processus, l'un des défis consiste à gérer les jeux de données avec un chevauchement limité en termes de propriétés utilisées pour décrire leurs ressources, ce que nous appelons des *jeux de données complémentaires*. Cette information manquante fait qu'il est difficile pour les systèmes récents basés uniquement sur l'analyse des propriétés [95, 155] d'évaluer les relations entre instances. Les jeux de données intégrés dans AgroLD présentent largement ce problème. Par exemple, dans la figure 5.4 sont représentées deux entités qui ont des URIs différentes. Il n'est donc pas possible de déterminer si elles sont identiques. Pourtant, un expert biologiste confirme leur similitude en se basant sur leurs propriétés *agrold :description* et *agrold :has_rapdb_identifier*. En effet, nous retrouvons bien la présence de l'identifiant de la ressource *rapdb*, associé à la deuxième entité, dans la description de la première entité.

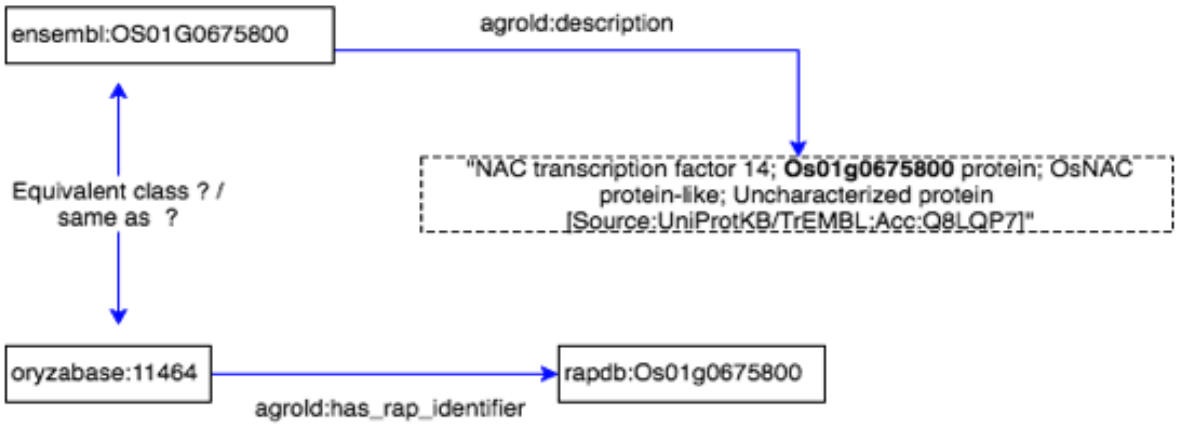


FIGURE 5.4 – Un exemple de problème de liage de données rencontré dans AgroLD

L'exemple de la figure 5.5 montre 2 entités biologiques issues de 2 jeux de données différents. Ces entités correspondent à la protéine APO1 mais elles sont considérées comme différentes car elles n'ont pas le même URI. De plus la tâche est d'autant plus difficile lorsque les propriétés qui les décrivent sont hétérogènes. Une des questions est d'identifier les propriétés sur lesquelles se baser pour faire la comparaison. Mais également de déterminer comment les attributs sont valués ou structurés afin d'éviter de produire des liaisons erronées ou de manquer des liaisons. Comme le montre la figure 5.5, les descriptions peuvent être exprimées dans différentes langues naturelles, avec différents vocabulaires ou avec différentes valeurs.

Ces limitations peuvent être classées selon 3 dimensions répertoriées dans la figure : dimension littérale, dimension ontologique et dimension logique.

La dimension littérale fait référence aux propriétés contenant des valeurs littérales (texte) exprimées en langage naturel ou valeurs numériques qui peuvent induire des erreurs de liage. Les

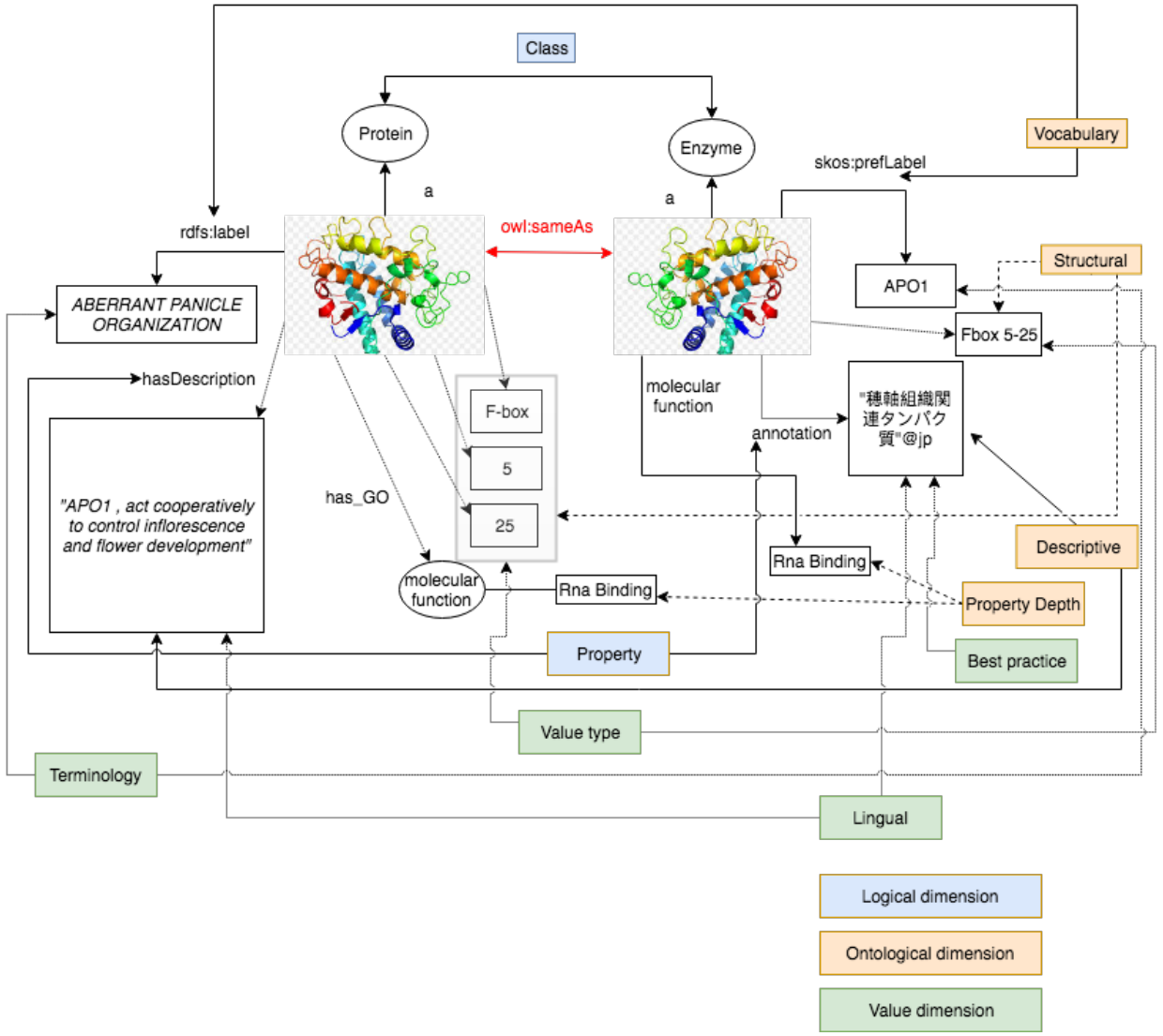


FIGURE 5.5 – Schéma général d'un exemple de liage de données inspiré de [4]

auteurs de Achichi et al, 2018 [4] identifient 4 niveaux d'hétérogénéité dans cette dimension, également indiqués dans la figure : type, terminologie, linguistique, bonnes pratiques de représentation.

— *Hétérogénéité liée aux types des valeurs.* Cette hétérogénéité concerne la manière dont les valeurs littérale sont encodées (e.g. string, integer, etc.). Dans ce cas, le challenge réside dans l'uniformisation des types de valeurs, par exemple uniformiser les formats des dates, des mesures numériques, etc.

— *Hétérogénéité liée à la terminologie.* Dans ce cas les variations vont concerner un terme correspondant à un mot ou un groupe de mots. Cette variation peut s'exprimer de différentes manières : i) la synonymie lorsque des termes différents vont représenter le même concept ; ii) la polysémie lorsque les termes similaires ont des sens différents ; iii) des acronymes et abréviations. Comme on peut le constater sur la figure 5.5 un des noms des entités correspond à une abréviation. Pour pallier ce problème, certaines applications proposent des

fonctionnalités d'expansion d'acronymes/abréviations.

— *Hétérogénéité liée à la linguistique.* Les termes concernés sont issus de langages différents. C'est un problème fréquent lorsqu'on travaille avec des données expérimentales de diverses origines et qui reflètent la diversité des informations que l'on peut trouver sur le Web. Dans ce cas, par exemple s'agissant de l'Anglais et du Japonais, les outils de recherche par similarité sont inefficaces. Il faut passer par une étape de traduction automatique au préalable.

— *Hétérogénéité liée aux Bonnes pratiques de représentation.* La représentation des connaissances est soumise à des bonnes pratiques de conception. Leur transgression est un frein dans la découverte de correspondances.

La dimension ontologique fait référence aux variations de classes ou de propriétés associées aux instances comparées. Quatre niveaux d'hétérogénéité sont identifiés provenant : du vocabulaire, de la structure, de la profondeur de niveau des propriétés et des descriptions.

— *Hétérogénéité du vocabulaire.* Les classes et les propriétés sont souvent décrites, par différents producteurs de données, en utilisant différents vocabulaires car la sémantique d'une classe ou d'une propriété donnée peut être interprétée différemment selon son application. Ce problème est encore plus compliqué dans le contexte du Web de données où toutes les ressources ne sont pas nécessairement décrites de la même manière. L'utilisation de mapping entre vocabulaires, par exemple avec LOV ou Agroportal dans notre cas, peut permettre de dépasser ce problème.

— *Hétérogénéité de la structure.* La description d'une entité peut se faire à différents niveaux de granularité. Dans notre exemple le terme *Fbox 5-25* est structuré différemment dans les deux entités : l'information est incluse dans une structure de données pour la première entité et dans un littéral pour la deuxième. L'utilisation de méthodes NLP pour extraire de l'information sur la deuxième entité peut aider pour le liage.

— *Hétérogénéité de la profondeur de niveau des propriétés.* Elle se situe au niveau du schéma des ressources et correspond à des différences de modélisation des propriétés. Dans notre cas, le littéral *DNA Binding*, qui est une fonction moléculaire, est modélisé à partir d'une classe de type GO pour la première entité et une propriété pour la deuxième. La distance entre les deux éléments est donc plus importante pour la première. Les méthodes pour résoudre ce type de problème, peuvent être d'indexer les littéraux avec leur contexte afin de pouvoir les comparer.

— *Hétérogénéité descriptive.* Une ressource peut avoir plusieurs concepts ou peut être décrite avec un ensemble de propriétés plus important dans un jeux de données que dans un autre, comme nous pouvons le voir dans notre exemple (voir la figure 5.5). On peut remarquer que ces ressources, et c'est le cas de manière générale, contiennent plus d'informations descriptives (des champs littéraux de type texte) que l'ensemble de propriétés qui les décrivent. Il est évident que comparer ces ressources uniquement par leurs propriétés sera moins efficace que des approches prenant en compte l'ensemble des informations.

La dimension logique fait référence au fait que l'équivalence entre deux informations sur deux jeux de données est implicite, mais peut être déduite à l'aide de méthodes de raisonnement. Deux

principaux problèmes d'hétérogénéité sont identifiés :

— *Hétérogénéité de classe.* Ce type d'hétérogénéité concerne le niveau de la hiérarchie des classes. C'est généralement le cas de deux ressources appartenant à des classes différentes pour lesquelles une relation hiérarchique explicite ou implicite est définie (les concepts «Protein» et «Enzyme», dans la figure 5.5, illustrent ce problème : la classe Enzyme est sous classe de Proteine). De plus, deux instances se rapportant au même objet peuvent appartenir à deux sous-classes différentes de la même classe.

— *Hétérogénéité de propriété.* A ce niveau, l'équivalence entre deux valeurs est déduite après l'exécution d'une tâche de raisonnement sur les propriétés. Deux ressources faisant référence à la même entité peuvent avoir deux propriétés qui sont inversées sémantiquement (c'est-à-dire les propriétés *hasDescription* et *isAnnotatedBy*). Dans ce cas, ces deux propriétés contiennent les mêmes informations, comme illustré dans l'exemple de la figure 5.5.

En ce qui concerne l'état de l'art nous avons répertorié les nombreux logiciels qui implémentent des méthodes de liage, dont les suivants sont les plus cités ou récents :

— **Silk** [95] met en œuvre des méthodes d'indexation et de présélection d'entités. La présélection consiste à rechercher un ensemble limité d'entités cibles susceptibles de correspondre à une entité source donnée. Toutes les ressources cibles sont indexées par une ou plusieurs valeurs de propriété spécifiée (le plus souvent, leur label). Le *rdfs :label* d'une ressource source est utilisée comme terme de recherche dans les index générés et seules les premières ressources cibles trouvées dans chaque index sont considérées comme des liaisons possibles pour la correspondance. Cette stratégie ne garantit pas la découverte de toutes les ressources équivalentes dans le jeu de données cible. Silk est basé sur les règles de liage définies par l'utilisateur (Silk-LSL). En d'autres termes, il comporte un langage déclaratif permettant de spécifier les types de liens RDF à découvrir entre les sources de données et les conditions que doivent remplir les entités pour pouvoir être inter-connectées.

— **Limes** [155] se configure à l'aide d'un langage de spécification permettant d'identifier les liens. LIMES (comme Silk) propose de l'apprentissage supervisé et de l'apprentissage actif pour la spécification des règles de liage. Pour cela, Silk et LIMES utilisent une programmation génétique. Cette dernière part d'un ensemble de spécifications de liens aléatoires et utilise les principes évolutifs de sélection et de variation pour faire évoluer ces spécifications jusqu'à ce qu'une condition de liaison réponde à un critère d'optimisation prédéfini (fonction fitness) ou qu'un nombre maximal d'itérations soit atteint. Pour l'apprentissage supervisé, des liens candidats validés manuellement sont utilisés dans l'algorithme génétique pour rechercher des liens similaires des règles de correspondance identifiées dans les données d'apprentissage. L'apprentissage actif vise à réduire la tâche de labellisation des données d'entraînement en mettant en œuvre un étiquetage interactif des candidats au liage sélectionnés automatiquement. Dans ce cas, les liens candidats sont sélectionnés pour optimiser la similarité avec des instances non étiquetées. Par ce moyen, LIMES partitionne l'espace métrique (d'instances) en représentant chacune de ces parties au moyen d'un exemple permettant de calculer une approximation précise de la distance entre instances sur la base de distances déjà connues. Grâce au gain considérable d'efficacité apporté par l'outil, LIMES est capable de relier de très grands ensembles de données, là où d'autres outils échouent.

— **Legato** [2] est conçu pour lier les entités de graphes ayant un haut degré d'hétérogénéité, se caractérisant par un faible recouvrement de ces ressources. L'outil est composé de modules qui s'enchaînent dans un workflow pour effectuer les différentes étapes nécessaires

au liage. Le module de nettoyage des données ne conserve que les propriétés comparables entre les jeux de données (par conséquent, les commentaires sous forme de texte libre, ainsi que les identifiants d'instances spécifiques à une ressource sont supprimés). Le module de profilage d'instance représente les instances par un sous-graphe correspondant à l'union du CBD (Concise Bouded Description) de chaque ressource et de ses voisins directs. En cela, contrairement à SILK ou Limes, Legato (dans sa version par défaut) ne compare pas les valeurs de propriétés, mais considère toutes les valeurs littérales extractibles comme un sac de mots. Cette représentation aborde dans son mécanisme un certain nombre d'hétérogénéités de données sans nécessiter l'intervention de l'utilisateur, en particulier les différences de description et les différences de profondeur de propriété décrites ci-dessus. Les littéraux de ces sous-graphes sont ensuite utilisés pour projeter chaque instance dans un espace vectoriel et la mise en correspondance consiste à comparer les vecteurs résultants. Un seuil délibérément bas est utilisé pour la similarité vectorielle afin d'assurer un rappel élevé. Ensuite, des instances très similaires sont regroupées à l'aide d'un algorithme de classification hiérarchique standard. Un algorithme découverte de clé RDF [202] et un algorithme de classement de clé [3] sont appliqués sur chaque paire de clusters similaires sur les deux graphes, afin d'identifier le jeu de propriétés qui permet le mieux de discriminer les ressources contenues dans chaque cluster. Un nouveau jeu de liens (appelé "liens sûrs") résulte de ce processus et est ensuite comparé aux liens produits à l'étape de mise en correspondance (appelés "liens candidats") afin d'éliminer les erreurs et d'augmenter la précision, aboutissant à la production du jeu de liens final. Le résultat de Legato est présenté au format EDOAL[7], ce qui permet de garder une trace des indices de confiance associés, ou sous forme de triplets *owl :sameAs*.

Le liage de données est un composant très important dans le processus d'intégration de données car il permet d'agréger plusieurs propriétés/annotations autour d'une même entité enrichissant donc ses informations. Peu de méthodes ont été développées sur des données réelles et aucune dans le domaine agronomique. En collaboration avec Konstantin Todorov (MdC, Lirmm) nous proposerons d'évaluer les outils issus de l'état de l'art cités précédemment et tenterons de développer une méthode adaptée au contexte d'AgroLD.

Nous proposons trois directions de recherche - éventuellement combinées - pour résoudre ce problème :

— **Extraction de données non structurées** : les graphes RDF contiennent du contenu textuel riche tel que des labels, des commentaires ou des descriptions qui fournissent une bonne information contextuelle. Ces contenus contiennent des entités et des relations susceptibles de compléter l'information nécessaire à la création de liens entre les jeux de données non-liés. Nous exploiterons ce contenu textuel en utilisant des techniques de traitement du langage naturel et d'extraction de relations pour identifier les entités nommées et reconstruire leurs relations, permettant ainsi la découverte de liens pertinents entre des ressources connexes.

— Les **techniques d'augmentation de graphe de connaissances** ajoutent des informations structurées aux graphes RDF existants en explorant des données externes pertinentes sur le Web (e.g. données de balisage, articles scientifiques, médias (sociaux), autres graphes de connaissances). Ce processus est particulièrement efficace pour récupérer des relations manquantes entre entités déjà présentes dans un graphe de connaissances. Nous appliquerons ces méthodes pour augmenter nos jeux de données en entrée et reconstruire les informations manquantes.

— **Apprentissage automatique pour des jeux de données complémentaires**. Nous allons explorer les critères pertinents qui représentent de manière effective les ressources inter-graphes

7. `http://alignapi.gforge.inria.fr/edoal.html`

et nous les classerons comme identiques (ou non) par apprentissage automatique. Nous utiliserons des modèles vectoriels pour des paires d'instances et ferons de l'apprentissage sur les relations entrée-sortie à partir des données d'apprentissage. Un jeu de données d'entraînement sur les données AgroLD est actuellement en construction.

Cette année un sujet de thèse sera proposée au concours de l'école doctorale de l'Université Montpellier.

5.2.3 Raisonnement sur les données

Le modèle de représentation des données en graphe RDF qu'utilise la plateforme AgroLD s'accompagne également d'autres langages structurants pour décrire les schémas de données (RDFS, OWL et SKOS) ou encore décrire les contraintes sur les données (ShEx - Shape Expressions, SHACL - Shapes Constraint Language). Le fait d'utiliser des ontologies avec les données permet de les structurer sous la forme de classes d'entités, de relations et d'instances.

Il est possible de mettre en œuvre des mécanismes de raisonnements grâce aux ontologies. Par exemple, les relations de généralisation/spécialisation sont très souvent utilisées dans les raisonnements pour propager de l'information. Dans ce cas, si l'on définit *Class B subClassOf Class A*, si l'entité *E1* est instance de *B* alors elle sera aussi instance de *A*.

Il est également possible d'utiliser le raisonnement pour enrichir les liens existants dans les données. C'est le cas lorsque on utilise les relations de réflexivité et de transitivité. Par exemple, dans le cas de données d'interactions moléculaires comme les réseaux d'interaction protéine-protéine ou de co-expression de gènes, le fait de définir la relation *interact_with* comme réflexive ou *coexpress_with* comme transitive permet de raisonner dessus afin d'enrichir l'information lorsque les données sont incomplètes.

Un autre aspect proposé par les technologies du Web sémantique est l'utilisation de langages à base de règles permettant de valider des contraintes sur les données. Parmi eux, citons les langages émergents SPIN - SPARQL Inferencing Notation [8], ShEx - Shape Expressions [9] et SHACL - Shapes Constraint Language [10]. Pour mener à bien cette problématique, j'évaluerai les possibilités que proposent ces langages pour implémenter des règles d'enrichissement ou de vérification de cohérences dans les graphes.

Peu de méthodes et outils ont été développés sur des données réelles et dans le domaine agronomique. Nous développerons des méthodes permettant de créer de nouvelles données à partir du raisonnement sur les schémas et vérifierons la validité des graphes.

Par ailleurs, dans le cadre du projet ANR D2KAB, nous démarrons une collaboration avec l'équipe Inria WIMMICS qui propose de nombreux outils dans ce domaine, notamment le moteur Corèse [44] qui intègre déjà certains langages mentionnés précédemment.

8. `http://spinrdf.org/spin.html`
9. `https://shex.io`
10. `https://www.w3.org/TR/shacl`

5.3 Applications sur les graphes de connaissances

5.3.1 Priorisation de gènes candidats

Cette dernière phase du projet, intervient après l'intégration de nombreuses sources de données, la création et l'enrichissement de ces dernières sous forme de graphes de connaissance. La recherche d'information parmi ces graphes nécessite le développement de méthodes pour trier pertinemment les résultats. La priorisation de gènes candidats permet d'identifier et de classer parmi un grand nombre de gènes, ceux qui sont fortement associés au phénotype ou la maladie étudiée.

Un certain nombre de méthodes informatiques ont été développées pour résoudre le problème de la priorisation des gènes associés à une maladie ou un phénotype [151]. Par exemple, Endeavour [151, 213] a pu associer le gène GATA4 à une hernie diaphragmatique congénitale ; GeneDistiller [185] a découvert le rôle des mutations MED17 dans l'atrophie cérébrale et cérébelleuse infantile. En se basant sur les approches informatiques sous-jacentes, les méthodes de priorisation des gènes peuvent être classées selon cinq types.

Les Méthodes de priorisation

— Le premier type concerne les **méthodes de filtrage**, qui passent au crible la liste des gènes candidats pour en réduire la taille en fonction des propriétés que les gènes associés devraient avoir [27, 150, 50].

— Le second type de **méthodes est basé sur la fouille de texte** [60, 195, 196]. En général, ces méthodes évaluent les gènes candidats en utilisant les preuves de co-occurrence avec une certaine maladie identifiée dans la littérature. L'inconvénient est que ces méthodes ne peuvent détecter que les associations déjà connues.

— Le troisième type est l'**analyse de similitudes et les méthodes de fusion de données** [5, 214, 34, 126, 63, 246, 237, 109]. C'est aujourd'hui le type de méthode le plus répandu dans la communauté de priorisation des gènes candidats et compte dans ses rangs la célèbre méthode Endeavour [5]. Ces méthodes reposent sur l'idée que des gènes similaires devraient être associés à des ensembles de phénotypes ou maladies similaires et inversement. La mesure de la similarité peut être définie à l'aide de différentes sources de données, telles que la Gene Ontology (GO) ou les résultats de score du logiciel BLAST (Basic Local Alignment Search Tool). Après avoir obtenu les scores de similarité pour chaque source de données, ces méthodes appliquent la fusion de données pour agréger ces scores dans un classement global.

— Le quatrième type concerne les **méthodes basées sur la construction de réseaux** [236, 120, 125, 117, 118, 101, 194, 173]. Ces méthodes représentent les phénotypes et les gènes comme des nœuds dans un réseau hétérogène, dans lequel le poids des arêtes représente leurs similarités. Le dernier type est basé sur les techniques de complétion de matrice dans les systèmes de recommandation [152, 237]. Ces méthodes représentent l'association des gènes et le phénotype comme une matrice incomplète et résolvent le problème de priorisation des gènes en remplissant les valeurs manquantes de la matrice. Ce type de méthodes s'est avéré être le plus performant à ce jour [237].

Comment est effectuée la recherche d'information à partir des mots clés proposés par l'utilisateur ? La plus communément utilisée, est l'approche « guilt by association » qui assume que les gènes associés ou interagissant dans un même processus partagent les mêmes fonctions. Les méthodes développées à partir de cette approche recherchent les mots clés présents dans les sources

de données et parmi un petit groupe de gènes annotés manuellement. La liste de gènes ainsi identifiés constitue une graine « seed genes » qui est ensuite utilisée pour trouver des associations avec les gènes à prioriser.

Malgré les progrès importants fait par les approches existantes, de nombreux verrous existent encore. Premièrement, les méthodes basées sur la similarité, qui reposent sur le principe de la «guilt by association» , échouent souvent dans le traitement de nouvelles maladies dont les gènes associés sont complètement inconnus [237]. Deuxièmement, bien que la performance des méthodes basées sur des réseaux soit acceptable, elles peuvent être biaisées par la topologie du réseau et intègrent difficilement plusieurs sources d'informations sur les gènes et les phénotypes [151]. En outre, la plupart des méthodes existantes reposent largement sur des fonctionnalités conçues manuellement ou sur des règles de fusion de données prédéfinies. Par conséquent, la problématique de priorisation reste encore ouverte.

Je compte approfondir ces aspects afin de proposer une méthode qui inclurait i) le calcul de scores pour chaque co-occurrence gène-phénotype trouvé dans une source (i.e. une source = un graphe), ii) la combinaison des différents résultats trouvés pour chaque source en pondérant les scores en fonction de l'origine des sources (i.e. source annotée manuellement, publication, etc.). Je continuerai à enrichir AgroLD en nouvelles connaissances et à implémenter ces nouvelles méthodes dans l'interface de recherche.

5.3.2 Analyse fonctionnelle des réseaux d'interactions moléculaires

Le succès récent des modèles de graphes et de l'apprentissage profond (*deep learning*) en bioinformatique [245, 127, 47, 104, 234] suggère la possibilité d'incorporer systématiquement de multiples sources d'information dans le réseau hétérogène d'interactions moléculaires et d'apprendre la relation non linéaire entre les phénotypes et les gènes candidats. Les graphes sont des outils très utiles et puissants pour représenter les interactions entre toutes les entités. Ainsi, ils sont parfaits pour représenter chaque type d'interactions qui se produisent dans les réseaux biologiques.

Récemment, de nouvelles approches combinant graphes de connaissances et apprentissage profond, ont été proposées dans le domaine du Web sémantique [177, 38, 108, 156]. AgroLD étant basé sur RDF, un modèle de représentation multi-graphes orientés étiquetés, la plateforme sera adaptée pour évaluer ce type d'approche. Toutefois, il existe peu d'outils et d'algorithmes capables de gérer l'analyse sur de large réseaux.

Nous proposerons de développer une approche adaptée au contexte d'AgroLD qui prend en considération les défis soulignés ci-dessus.

— **Évaluation des méthodes de plongements de graphes** La classification et le clustering dans l'espace des graphes manque de méthodes appropriées, une technique pour pallier ceci consiste à transformer le graphe en vecteur, on parle alors de plongements de graphes. Afin de pouvoir comparer l'information contenu dans plusieurs graphes et de mesurer leur similarité, nous comparerons plusieurs méthodes de plongements (i.e. *Graphs Neural Networks versus Networks Embeddings*) dans un même espace vectoriel (dans ce cas le graphe est transformé dans un vecteur numérique).
— **Évaluation des méthodes de mesure de similarités** : nous évaluerons plusieurs méthodes combinant des approches de réseaux de neurones afin d'apprendre à partir des données présentes et prédire des nouvelles interactions afin de pallier aux irrégularités dans les données. Ces mesures de similarités seront utilisées dans la priorisation des gènes candidats.

Chapitre 6

Conclusion et perspectives

6.1 Conclusion

Dans ce mémoire, nous avons présenté quelques défis scientifiques et techniques dans la représentation de la complexité des données biologiques afin d'en extraire de la connaissance permettant d'identifier les mécanismes moléculaires contrôlant l'expression de phénotypes chez les plantes. Nous avons montré que cette complexité des données biologiques soulève de multiples questions de recherche informatique dans des domaines variés tels que la représentation des connaissances, le Web sémantique, l'intégration des données, le traitement du langage naturel, etc. Nous avons illustré nos réflexions par les résultats obtenus au cours des 13 dernières années dans le cadre de nos projets en agronomie. Nous n'avons pas couvert tous les travaux liés aux défis cités en introduction et nous avons certainement omis d'autres défis importants tels que l'évaluation de la qualité et la provenance des données dans le processus d'intégration, la reproductibilité, l'évolution des annotations sémantiques avec les données, la visualisation et l'exploration des données. Mais nous avons montré un bref résumé des multiples contributions dans les domaines de l'intégration des données et la représentation des connaissances en biologie végétale.

Dans le cadre du projet Gigwa : Nous avons développé une application permettant de répondre à la problématique de passage à l'échelle dans le stockage et l'analyse de données génomique. Pour cela, nous avons su tirer avantage des caractéristiques du système d'information NoSQL MongoDB, qui est largement utilisé aujourd'hui, pour stocker et gérer les données de variations génomiques. Les résultats sont très encourageants. Tout d'abord, car nous avons pu montrer que Gigwa se présentait, à un large public, comme une alternative aux logiciels disponibles en ligne de commandes. En effet, à travers plusieurs publications, nous avons pu montrer que l'on pouvait concilier performance et facilité d'utilisation. Par ailleurs, nous l'avons développé en tenant compte de l'écosystème d'analyse bioinformatique existant, c'est-à-dire avec de nombreuses fonctionnalités d'import, d'export couplées à une API de communication.
Même si, pour l'instant, nous n'avons pas pu obtenir de financements sur projet, les avis des utilisateurs sont très positifs et nous encouragent à continuer d'améliorer l'application. Enfin, le fait que l'application puisse être utilisée à la fois sur une machine de bureau et sur un serveur dédié a contribué à sa popularité. Toutefois, avec son architecture actuelle, la gestion de très gros volumes de données >1To est un facteur limitant à son utilisation. De plus, cela a un impact négatif sur sa vitesse d'exécution. Nous envisageons, à l'avenir, de nous orienter vers différentes stratégies de déploiement. La première consoliderait l'architecture actuelle et l'optimiserait pour l'ordinateur de bureau et la seconde serait re-définie et optimisée pour de grandes infrastructures de calcul.

Dans le cadre du projet AgroLD : Nous avons développé une plateforme permettant de répondre aux problématiques d'intégration de données et de représentation de connaissances en

utilisant les technologies du Web sémantique. AgroLD cible essentiellement le domaine de la biologie des plantes, avec un intérêt pour la représentation et la découverte de relations génotype-phénotype. Depuis le début de ce projet nous avons tenté de répondre à certains défis comme celui de représenter la complexité du domaine biologique dans un formalisme informatique adapté. Nous avons évalué et développé des méthodes de transformation RDF sur de grands jeux de données et nous avons effectué une annotation sémantique. Nous avons également développé quelques premières méthodes pour la visualisation de données liées adaptées à différents profils d'utilisateurs.
Même si les concepts même du Web sémantique et des données liées sont difficiles à appréhender (surtout pour des données dans leur forme originale) pour des utilisateurs finaux, AgroLD a été bien accueillie par la communauté biologique. Toutefois, il reste encore de nombreux défis à relever avant que nous puissions en récolter les bénéfices. Nous poursuivrons nos travaux dans les directions de recherches sur le liage de données, l'extraction de connaissance dans le texte, le raisonnement sur les données et les systèmes de recommandation. Nous décrirons quelques perspectives de recherche dans la prochaine section ainsi que de nouvelles orientations sur le long terme.

6.2 Perspectives

L'objectif du projet de recherche (décrit en section 5) sera en effet, de développer des approches de priorisation de gènes candidats utilisant les réseaux d'interactions moléculaires comprenant des sources multiples issues de graphes de connaissances. Les méthodes de priorisation actuelles s'appuient essentiellement sur les méthodes d'apprentissage supervisé et non supervisé et n'utilisent que rarement les graphes de connaissances et le raisonnement sur les données. De plus, le domaine agronomique commence leur exploitation.

En ce qui concerne le domaine informatique, les axes de travail proposés s'inscrivent dans une démarche mutualiste : l'extraction et la publication de connaissances sur le Web de données. Les méthodes que nous souhaitons développer s'appuient sur des données réelles ou des plate-formes de production de données. Elles répondront donc à des besoins réels et espérons le, auront un impact important pour les communautés concernées.

Les résultats de ce projet sont directement dédiés aux scientifiques des domaines de la génétique et de l'amélioration des plantes car il existe actuellement un réel verrou dans la gestion et le traitement des données biologiques. C'est le cas de mon unité actuelle, DIADE, mais également de nombreuses unités plantes présentes sur Montpellier et plus largement au niveau national.

Certains enjeux soulevés par ce projet de recherche sont actuellement identifiés dans le cadre du projet **ANR D2KAB** (*Data to Knowledge in Agronomy and Biodiversity*) porté par le LIRMM et démarré en juin 2019 [1]. Centré autour d'AgroPortal, D2KAB ambitionne de créer un cadre permettant de transformer les données agronomiques et de biodiversité en connaissances interopérables, exploitables et ouvertes, ainsi que d'étudier les méthodes et outils scientifiques permettant d'exploiter ces connaissances pour des applications dans les domaines de la science et de l'agriculture. Le projet D2KAB réunit un consortium multidisciplinaire de 11 partenaires : 2 unités de recherche en informatique (LIRMM, I3S) ; 5 unités de recherche en informatique appliquée INRA/IRSTEA (URGI, MaIAGE, IATE, DIST, TSCF) spécialisées en agronomie ou en agriculture ; 2 laboratoires de recherche sur la biodiversité et les écosystèmes (CEFE, URFM) ; 1 association d'acteurs de l'agriculture (ACTA) ; et 1 partenariat avec le département de Stanford du BMIR. Le projet comprend

1. http://www.d2kab.org

5 scénarios du domaine agronomique (les emballages alimentaires, les données liées à l'agro-agri, les phénotypes du blé, les écosystèmes végétaux et la biogéographie) qui produiront des résultats concrets pour les communautés scientifiques et les acteurs socio-économiques de l'agriculture.

Mon projet de recherche, centré sur la plateforme **AgroLD**, a également été soumis pour un appel à financement de projets de recherche de l'**I-SITE MUSE** de Montpellier. Son objectif est de (i) fournir l'accès à une plateforme de référence pérenne répondant aux principes FAIR tout en respectant les conditions éthiques du partage de données biologiques/génétiques, (ii) valoriser les données expérimentales pour les unités de l'I-SITE MUSE et ses partenaires. Il fédère 6 unités de recherches de Montpellier (AGAP, DIADE, IPME, LIRMM, MISTEA, LEPSE, SHS-GRED) et 5 partenaires extérieurs (URGI, Wageningen Univ., IRRI, Bioversity, USTH-Vietnam). Le projet implique une vingtaine de scientifiques du CIRAD, de l'IRD, de l'INRAE, et de l'UM, dont certains ont déjà collaboré sur des projets de recherche structurants ou d'infrastructures de recherche (PIA IBC, plateforme bioinformatique South Green, ANR D2KAB). Une attention particulière sera portée aux données expérimentales produites par les chercheurs des unités plantes Montpelliéraines (DIADE, AGAP, IPME, LEPSE).

Les méthodes qui seront mises en place dans ce projet de recherche devraient également avoir des retombées intéressantes au niveau des collaborations potentielles avec des centres internationaux comme l'IRRI (le centre international du riz basé aux Philippines) dans le cadre du workpackage *Big Data integration platform* du projet international Rice CRP. Il trouve également une place dans l'initiative Européenne H2020 Elixir-Exelerate sur le développement d'une infrastructure Bioinformatique pour l'échange de données et de services sur le Web (collaboration CIRAD-IRD avec Wageningen UR et INRAE-URGI).

Dans un avenir plus éloigné, je poursuivrai mes efforts pour relever les défis identifiés (et ceux qui émergeront), tout en continuant à offrir à une plus large communauté scientifique les moyens de partager et d'exploiter leurs ressources sémantiques et de permettre l'émergence de nouvelles sciences dans leurs domaines. L'évolution de mes activités pourrait s'étendre à d'autres domaines de la biologie végétale voire même humaine ou animale. Par exemple, les domaines de la biodiversité végétale et animale sont des thématiques de recherches portées et soutenues par l'IRD. Le domaine de la santé humaine est également porté par l'institut avec un focus particulier sur les zones Afrique et Asie qui sont souvent le foyer d'épidémie. La poursuite de mes travaux s'articulera autour de trois objectifs généraux :

1. Encourager l'adoption de l'ingénierie des connaissances : J'aimerai continuer à utiliser l'ingénierie de la connaissance comme dénominateur commun de l'interopérabilité et de l'intégration des données ; pouvoir aider les différentes communautés scientifiques à développer de nouvelles ressources sémantiques et les encourager à adopter les standards du Web sémantique pour structurer leurs connaissances ; concevoir, développer et maintenir des outils permettant de produire, de diffuser, de partager et d'interconnecter leurs ressources sémantiques.
2. Contribuer à la production de données FAIR et à l'exploitation des données liées : Une perspective est de développer des méthodes et des outils pour transformer les données en connaissances qui peuvent être utilisées par les machines pour la recherche d'information et le raisonnement sur les données. En m'appuyant sur mes expériences de projets passés, je pourrai développer de nouvelles méthodes. Par exemple, dans le domaine de l'agriculture avec les données issues de capteurs ou de géolocalisation ou en santé avec les données patients. En matière d'exploitation, des perspectives peuvent être liées au raisonnement sur les données, aux méthodes de liage et de visualisation.

3. Faciliter une science des données sémantiquement enrichies : Le développement de méthodes permettant l'intégration de données et l'extraction de connaissances à pour objectif de lever des verrous informatiques dans les domaines d'applications concernés. C'est également le moyen de créer de la connaissance et de faciliter des découvertes scientifiques. C'est ce que je réaliserai à travers les scénarios de priorisation de gènes candidats ou d'étude des relations génotype-phénotype. En fonction des collaborations futures, je mettrai en œuvre de nouveaux scénarios à l'interface biologie-informatique.

Troisième partie

Annexes

Chapitre 7

Curriculum Vitae

7.1 Identité

Pierre LARMANDE
47 ans, né le 11 Octobre 1972, à Perpignan
Français, marié, 2 enfants nés en 2002 et 2009
95 Ve Ho, Xuan La, Tay Ho,
Hanoi, Vietnam
+33 6 50 14 90 41
+84 36 60 18 725
Page Perso [1]

IRD - UMR DIADE	Associé au LIRMM
ICT Lab & LMI RICE USTH	équipe FADO
Hanoi, Vietnam	Montpellier, France
pierre.larmande@ird.fr	pierre.larmande@lirmm.fr

7.2 Formation

Licence de Biochimie - Maîtrise de Biochimie Université Montpellier 2 :
Obtenues en 1995 -1996

D.E.S.S. Informatique I.A.O Montpellier 2 :
Spécialité base de données et programmation. Obtenu le 2 septembre 2000 à Montpellier, mention assez bien

Doctorat de 3me cycle Université Montpellier 2 :
Soutenue le 20 décembre 2007 à Montpellier, mention très honorable

Sujet : Mutualiser et partager, un défi pour la génomique fonctionnelle végétale

Président : Corinne Cauvet, Professeur d'Université Aix-Marseille 03

Rapporteurs : Anne Doucet, Professeur, Université Paris 6 et
Christine Froidevaux, Professeur Université Orsay (Paris)

Co-encadrants : Isabelle Mougenot, Maître de conférence Université Montpellier 2 et
Manuel Ruiz, Chercheur Cirad (Montpellier)

Directeurs : Thérèse Libourel, Professeur Université Montpellier 2

7.3 Expérience professionnelle

Mon expérience professionnelle alterne des contrats d'ingénieurs puis de chercheurs et s'étale sur plus de 19 ans.

Novembre 2018 - En cours (1,5 an) Chercheur, CRCN,
Affectation : IRD, UMR DIADE équipe RICE (Rice, Interspecies Comparison & Evolution)
Scientifique associé équipe FADO (LIRMM)
En affectation au Vietnam (Hanoi)
Co-Directeur du laboratoire d'informatique ICTLab de l'USTH avec le Pr. Luong Chi Mai.

Septembre 2016 - Octobre 2018 (2 ans) Bioinformaticien, IE1,
Affectation : IRD, UMR DIADE équipe RICE (Rice, Interspecies Comparison & Evolution)
Scientifique associé équipe FADO (LIRMM)
En affectation au Vietnam (Hanoi).
Co-Directeur du laboratoire d'informatique ICTLab de l'USTH avec le Pr. Luong Chi Mai.

Octobre 2010 - Août 2016 (6 ans) Bioinformaticien, IE2,
Affectation : IRD, UMR DIADE équipe RICE (Rice, Interspecies Comparison & Evolution)
Financement : Ingénieur d'Etude (IE2) permanent IRD

Décembre 2005 - Septembre 2010 (5 ans) Bioinformaticien, IE2,
Affectation : CNRS, CEFE mis à disposition au CIRAD UMR AGAP
Financement : Ingénieur d'Etude permanent CNRS

Août 2002 – Novembre 2005 (3,3 ans) Bioinformaticien, Ingénieur
Affectation : CIRAD, UMR PIA, Montpellier
Financement : CDD Ingénieur d'Etude CNRS

Février 2001 – Juillet 2002 (1,5 an) Bioinformaticien,
Affectation : CIRAD, UMR PIA, Montpellier
Financement : le projet Génoplante / CDD CIRAD

Domaines de recherche : Génomique fonctionnelle, Agronomie, Ontologies et vocabulaires du domaine biologique, Intégration de données, Base de données, Visualisation de données, Recherche d'information, Fouille de texte, Représentation des connaissances, Web Sémantique, Données ouvertes et liées, Raisonnement sur les données, Architecture orientées service, WEB 2.0.

7.4 Éléments marquants du CV

— **Recherche multidisciplinaire** : Biologie moléculaire, Agronomie, Biologie computationnelle, Ontologies, Web Sémantique, Annotation sémantique, Fouille de texte, Architecture orientée services, Représentation des connaissances en biologie.

— **Recherche appliquée** dans le domaine agronomique, développement logiciel et transfert technologique.

— **Recherche collaborative** nationale et internationale (projets CGIAR, ANR, PIA), expérience en gestion de projets (ANR PIA), encadrements (masters, PhD et Postdoc) et gestion d'équipe.

— **Expérience de mobilité à l'étranger :** 4 ans au Vietnam - ICTLab USTH.

7.5 Projets scientifiques

— **2019-2022** Projet Flash science ouverte Participation française au GO FAIR Food Systems Implementation Network - FOOSIN [2] - 75K€ (Porteur S. Aubin)
Le projet est d'envergure regroupe 9 partenaires et comporte 5 WP. Je suis responsable du budget pour les unités IRD et associées. Je serai également responsable de l'implémentation des principes FAIR dans la base AgroLD.

— **2019-2022** Projet ANR PRCE Data to Knowledge in Agriculture and Biodiversity - D2KAB [3] - 950K€ (Porteur C. Jonquet)
Le projet est d'envergure, regroupe 11 partenaires et comporte 7 WP découpés en 3 tâches. Je suis impliqué dans 3 tâches : Intégration de données dans AgroLD (WP4) et production des benchmarks pour les tâches de lifting et liage de données (WP3).

— **2017-2022** Projet international CGIAR – CRP-RICE [4] (Porteur B. Bouman) - 1,5 M€ pour l'IRD (responsable A. Guesquiere)
Le projet regroupe 7 partenaires répartis en 7 WP. Je suis co-responsable de la tache WP4.5 sur la mise en œuvre de méthodes pour gérer des «Big Data ». Il s'agit d'assurer le suivi des délivrables et de participer à une partie d'implémentation dans l'intégration de données et l'interopérabilité des systèmes d'information.

— **2014-2018** Projet ANR Investissement d'Avenir IFB plant node (Institut Francais de Bioinformatique) [5]. Développement d'un réseau de ressources bioinformatiques sémantiquement interconnectées. 400 K€. (Porteur M. Ruiz et H. Quesneville).
Le projet était un appel à financement spécifique du projet IFB. Il rassemblait 5 instituts partenaires répartis en 3 WP. j'ai été responsable de tâche dans le WP1 en supervisant l'intégration des jeux de données INRA URGI dans AgroLD puis j'ai contribué à la mise en place de l'indexation d'AgroLD dans le portail SolR URGI du WP2.

— **2012-2017** Projet ANR Investissement d'Avenir IBC «Institut de Biologie Computationnelle »Modélisation, traitement et analyse des données à grande échelle en biologie, santé, agronomie et environnement - 2.842 M€ (Porteur O. Gascuel)(Co-responsables du WP5 P. Larmande et P. Valduriez)
J'ai initié et contribué au développement de plusieurs projets notamment AgroPortal, AgroLD et Gigwa.

— **2015-2016** Projet Lingua dans le contexte Labex Numev - 75 K€ (Porteur C. Jonquet).
Collaboration avec un post-doc sur l'annotation de mappings ontologies-données.

— **2014-2015** Projet BIOeSAI dans le contexte de financements incitatifs IRD. Développement d'une base de données phénotypique chez le riz - 11 K€ (Porteur. P. Larmande).
J'ai conçu le projet et encadré les étudiants pour la réalisation.

— **2015-2016** Projet LandPan TOGGLE dans le contexte du Labex Numev - 50K€ (Porteur. P. Larmande).

2. https://anr.fr/fr/lanr-et-la-recherche/engagements-et-valeurs/la-science-ouverte/les-projets-laureats-de-lappel-flash-science-ouverte/projet-foosin
3. http://d2kab.mystrikingly.com/
4. https://www.cgiar.org/research/program-platform/rice
5. https://www.france-bioinformatique.fr

Développement d'un pipeline d'analyse de données génomique. J'ai conçu le projet et co-supervisé un ingénieur

— **2015-2016** Projet AgroPortal dans le contexte du Labex Numev - 50 K€ (co-porteur C. Jonquet et P. Larmande).
Mise en place de l'entrepôt dans un environnement de production. J'ai participé à la conception du projet.

— **2013-2017** Projet ANR Bioadapt : Africrop «Documenting African Crop Domestication »- 698 K euros (Porteur Y. Vigouroux).
Le projet rassemblait 6 partenaires. J'ai été responsable de la tâche d'intégration des données produites dans Gigwa.

— **2004-2008** Projet CGIAR «Generation Challenge Programme (GCP) Pantheon ». Construction d'une architecture de médiation pour ressources distribuées - 400 K€ pour le Cirad (Porteur M. Ruiz).
Le projet rassemblait 11 partenaires. J'ai été responsable de la tâche d'intégration des métadonnées des services Web dans le registre de métadonnées.

7.6 Prototypage

Durant ma carrière scientifique j'ai eu l'occasion de développer ou de contribuer au développement de nombreux logiciels, base de données et applications Web. Une sélection des plus stables est listée ci-dessous.

— **PyRice** [6] - API de recherche d'information distribuée pour le riz et analyses de données en Python.

— **AgroLD ETL** [7] - API de transformation de données en RDF pour les ressources intégrées dans AgroLD.

— **AgroLD** [8] – Application web de visualisation des données sous forme de graphes RDF, constructeur de requêtes, API de services web, pipeline de transformation RDF, construction des modèles et de l'ontologie.

— **GIGWA** [9] – Développement d'une base de données génomiques, constructeur de requêtes, API de services web.

— **BioSemantic** [10] - Développement d'un constructeur de requêtes SPARQL au-dessus de bases de données relationnelles biologiques.

— **OryGenesDB** [11] - Développement d'une base de données génomique pour le riz.

— **Oryza Tag line** [12] - Développement d'une base de données de mutant phénotypiques pour le riz.

— Détection de motifs «FST »dans les séquences nucléiques du genome Oryza Sativa (Riz)

6. https://github.com/SouthGreenPlatform/PyRice
7. https://github.com/SouthGreenPlatform/AgroLD_ETL
8. http://www.agrold.org
9. http://southgreen.fr/content/gigwa
10. http://www.southgreen.fr/content/biosemantic-tool
11. http://orygenesdb.cirad.fr
12. http://oryzatagline.cirad.fr

7.7 Animation de la recherche

Responsabilité d'équipe

Co-Direction ICT Lab USTH - Hanoi

Depuis 2017, je suis co-directeur avec Pr. Luong Chi Mai, du laboratoire mixte IRD-USTH ICT Lab [13]. Il est composé de 11 chercheurs et enseignants chercheurs. Cette fonction recouvre l'animation scientifique, la communication, la gestion du budget, la rédaction des rapports d'activité, Les interactions et collaborations au sein du laboratoire me permettent de développer certains aspects de mon projet de recherche. Par exemple, dans le domaine de l'intelligence artificielle et de l'apprentissage machine que j'applique au domaine biologique.

Co-responsable du WP5 de l'Institut de Biologie Computationnelle - IBC

Entre mi-2013 et début 2017, j'ai été coordinateur de l'axe wp5 « intégration des données et connaissances biologiques » d'IBC [14].

Cet axe dont l'objectif est de faciliter l'accès aux données et connaissances en biologie était composé de 10 chercheurs et ingénieurs issus de plusieurs projets. Mon rôle consistait en particulier à assurer le suivi des avancements et des livrables, la gestion du budget, les rapports d'activité, la communication. Dans ce contexte j'ai supervisé les travaux d'un post-doctorant, d'un ingénieur et de stagiaires sur différents livrables. Sur le plan personnel, j'ai pu développer de nouvelles méthodes d'intégration sur des données expérimentales.

Co-responsable du plateau bioinformatique *i-Trop* IRD

Le plateau *i-Trop* [15] est une infrastructure de calcul et de services mise en place et maintenue par le centre IRD de Montpellier pour les unités locales et les partenaires du Sud. Ce plateau a pour fonction (i) de proposer un environnement de travail doté de capacité de calcul et de stockage adapté aux besoins des scientifiques, (ii) de centraliser les ressources bioinformatiques nécessaires pour ses utilisateurs. J'ai participé au montage et l'animation de cette structure, depuis janvier 2010, et j'en ai été le coordinateur de 2012 à 2013. Je reste actuellement contributeur en termes de services et applications.

Responsable et Membre de Comités d'Organisation

Semantic Web for Biodiversity (S4BIODIV) 2013

S4BIODIV [16] est un workshop attaché à la conférence ESWC2013. Montpellier, France Membre du comité d'organisation, j'ai notamment, rédigé la proposition de workshop, obtenu des financements, invité les keynotes, préparé et animé le panel discussion. Proceedings disponibles sur CEUR [17]

13. http://ictlab.usth.edu.vn
14. http://www.ibc-montpellier.fr
15. http://bioinfo.mpl.ird.fr
16. http://semantic-biodiversity.mpl.ird.fr
17. http://ceur-ws.org/Vol-979/

PhenoHarmonis : Harmonization, semantic and interoperability of phenotypic and agronomic data Workshop. Avril 2014 et 2016, Montpellier, France

Suite au succès du workshop S4BIODIV, les membres de cette organisation ont travaillé sur cette nouvelle série. Membre du comité d'organisation de PhenoHarmonIS [18] j'ai participé aux recherches de financements et aux proposition de sessions thématiques.

IC2016 [19] : 27e Journées francophones d'Ingénierie des Connaissances. 6-10 juin 2016, Montpellier, France

Membre du comité d'organisation de la conférence, j'ai participé aux recherches de financements dédiés à l'invitation de conférenciers.

AgroHackathon [20] : discovering AgroPortal & AgroLD. Juin 2016, Montpellier, France

Premier Hackathon dédié à l'intégration de données agronomiques. Co-organisateur avec C. Jonquet, nous avons effectué la recherche de financement, la préparation du Hackathon et l'organisation de plusieurs hacks autour d'AgroLD.

RDA Rice Data Interoperability Working Group

Research Data Alliance est une organisation internationale dont l'objectif est de promouvoir les standards d'échange et la publication FAIR des données dans la communauté scientifique. J'ai coordonné en 2017-2018, le groupe sur le riz composé d'une vingtaine de membres. L'objectif du Rice Data Interoperability WG [21] été de proposer l'utilisation de standards et un guide de bonnes pratiques pour échanger et publier les données produites sur le riz. Toutefois, les travaux de ce groupe n'ont pas pu aboutir faute de participation active de ses membres.

Comité de programme

— BioNLP Open Shared Tasks (1 fois en 2019)
— Réseau Intégration de sources/masses de données hétérogènes et ontologies (In-OVIVE) (1 fois en 2018)
— Data Integration for Life Science (DILS) (1 fois en 2017)
— 1st International Workshop on Semantics for Biodiversity (S4BioDiv)(1 fois en 2013)

JURYS

— Jury Masters BioPharma et ICT USTH 2017-2018 (2 fois)
— Jury de concours CNRS (Ingénieur d'Etude)(1 fois en 2008)
— Rapporteur de stages de M2 Bio-Informatique (UM) (8 fois depuis 2002)

Relecteur d'articles

— Nucleic Acids Research - database issue (8 reviews),
— Databases (3 reviews),
— Bioinformatics (2 reviews),

18. https://tinyurl.com/PhenoharmonIS2018
19. https://ic2016.sciencesconf.org
20. https://www.meetup.com/AgroHackathon
21. https://www.rd-alliance.org/groups/rice-data-interoperability-wg.html

— BMC Bioinformatics (3 reviews),
— Current Plant Biology (1 review)

Expertises

— Membre extérieur de comité d'évaluation des agents CIRAD (1 fois)
— Membre de comité d'évaluation des départements Bio et ICT USTH (2 fois)
— Membre de comité d'évaluation d'intelligence artificielle de l'ANR (1 fois)

7.8 Activités d'enseignement

les activités ont été effectuées soit en milieu universitaire soit en formation interne IRD

— 2018-2019 Enseignement dans le Module Web Sémantique du Master 1 ICT de l'Institut Francophone d'Informatique parcours 1 et 2, Hanoi,(60h - 2x30h)
— 2017-2018 Enseignement dans le Module Sensibilisation à la Bioinformatique du Master 2 ICT USTH, et Master 2 BioPharma USTH,(50h - 2x 25h)
— 2017, Module Systèmes d'information Géographique du Master 2 ICT USTH Hanoi, (25h)
— 2013 Module Bioinformatique du Master 2 BioPharma USTH Hanoi, (40h)
— 2011-2013 Formation interne IRD en Bioinformatique, (30h/an, 90h au total)
— 2004-2006 Enseignement TP Base de données IUT Informatique, Univ. Montpellier, (120h au total)
— 2002-2003 Enseignement TP BioPerl du DESS de Bioinformatique, Univ. Montpellier,(50h)

7.9 Encadrements

Doctorant

2009-2011 J. Wollbrett

Title : Génération semi-automatique de services Web sémantiques pour des bases de données relationnelles biologiques

— Thèse de l'Université Montpellier II
— Taux d'encadrement : 50% avec M.Ruiz et I. Mougenot
— Soutenance : Déc. 2011
— Situation actuelle : Post-Doctorant au Swiss Institute of Bioinformatics. Auparavant Post-doctorant au CNRS Roscoff.
— Financement : Bourse Région Languedoc Roussillon - CIRAD

Post-doctorant

J'ai collaboré avec 3 docteurs en stages post-doctoraux.

+ **[2015 - 2017 (2 ans)]** A. Toulet (Co-supervision avec Clément Jonquet)
  Contribution au développement d'un portail d'ontologies pour l'agronomie : AgroPortal.
  Financement Numev puis IBC

+ [2014 –2016 (2 ans)] : A. Venkatesan
Intégration de données utilisant les métadonnées et ontologies pour agréger les données de plusieurs ressources hétérogènes.
Financement IBC

+ [2012-2013 (2 ans)] J.Wollbrett
Automatiser l'intégration de bases de données relationnelles distribuées à travers l'enrichissement sémantique de vues RDF avec BioSemantic
Financement Cirad - IBC

Ingénieur de recherche

J'ai collaboré avec 2 ingénieurs de recherche.

+ [2015 – 2017 (2 ans)] N. El Hassouni
Contribution dans le développement du projet AgroLD.
Financement INRA sur le projet IFB

+ [2014 –2016 (2 ans)] G. Sempéré
Conception et développement de l'application Gigwa.
Financement Cirad.

Masters

J'ai encadré 21 stages en co-encadrement avec des biologistes ou des informaticiens.

Encadrement de M1

— **2019 M1 Recherche (6 mois)** Q. Do – Candidate gene prioritization using graph embedding – Hanoi University of Science and Technology (Hanoi)
Situation suivante : M2

— **2018 M1 Recherche (6 mois)** H. Do – Evaluating Name-Entity Recognition approaches in plant molecular biology – M1 USTH (Hanoi)
co-encadrement : K. Tanh
Situation suivante : M2 USTH

— **2017 M1 Professionnel (4 mois)** B. Vautrin – Développement de module ETL pour l'intégration de ressources dans AgroLD. PolyTech. (stage international - Hanoi)
Situation suivante : Dernière année polytech

— **2017 M1 Recherche (3 mois)** A. Diouf – Proposition et implémentation d'algorithmes de liage de données RDF dans AgroLD – M1 BCD (effectué à Montpellier)
co-encadrement : K. Todorov
Situation suivante : M2 BCD

— **2016 M1 Recherche (3 mois)** S. Zevio - Indexation de données issues du web sémantique dans le domaine agronomique – Master1 parcours DECOL UM (effectué à Montpellier)
Situation suivante : Master 2 DECOL UM

— **2014 M1 Recherche (3 mois)** L. Le Ngoc - Développement d'une application de gestion de données phénotypique chez le riz – Master1 Institut Francophone d'Informatique (effectué à Hanoi)
Situation suivante : Master 2 IFI

— **2014 M1 Recherche (3 mois)** G.Tagny - Développement d'une base de connaissances sur les gènes régulateurs de la ramification chez le riz – Master1 Institut Francophone d'Informatique (effectué à Hanoi)
Situation suivante : Master 2 IFI

Encadrement de M2

— **2019 M2 Recherche (5 mois)** S. Sonfac – Augmentation et liage de données par plongement multi-modal dans le contexte d'AgroLD – M2 IFI (effectué à Montpellier)
co-encadrement : K. Todorov
Situation suivante : Thèse

— **2018 M2 Recherche (5 mois)** K.M. Djibril – Developping an Ontology Matching workflow using AgroPortal API – M2 USTH (effectué à Hanoi)
Situation suivante : Job IT

— **2018 M2 Recherche (5 mois)** A. Sayadi – Liage de données complementaires dans le contexte d'AgroLD – M2 AIGLE (effectué à Montpellier)
co-encadrement : K. Todorov
Situation suivante : Recherche d'emploi

— **2016 M2 Professionnel (6 mois)** A. Petel – Contribution au développement de Gigwa – Master2 Polytech Grenoble (effectué à Montpellier)
co-encadrement : G. Sempéré
Situation suivante : Volontaire International Cirad, la Réunion

— **2016 M2 Recherche (5 mois)** D. Hyżorek - Epigenetic Data Integration and Analysis – Master 2 parcours BCD UM (effectué à Montpellier)
co-encadrement : M. Mirouze
Situation suivante : Chercheur en Pologne

— **2016 M2 Recherche (5 mois)** S. Remini - Acquisition automatique de connaissances à partir de textes scientifiques – Master2 parcours BCD UM (effectué à Montpellier)
co-encadrement : K. Todorov
Situation suivante : Recherche d'emploi

— **2015 M2 Recherche (6 mois)** G. Tagny (M2) - The Agronomic Linked Data (AgroLD) project. Master2 Institut Francophone d'Informatique (effectué à Montpellier)
Co-encadrement : A. Venkatesan
Situation suivante : Thèse Ecole des Mines Ales, Nîmes

— **2015 M2 Recherche (5 mois)** I. Chentli - Facilitation de l'accès aux données biologiques sémantiquement structurées – Master2 parcours BCD UM (effectué à Montpellier)
co-encadrement : K. Todorov
Situation suivante : Ingénieur Bioinformatique, IMGT-CNRS

— **2015 M2 Recherche (6 mois)** L. Le Ngoc (M2) - Développement d'un système connaissances pour BIG DATA : application aux données de phénotypage chez le riz (O. sativa) – Master2 Institut Francophone d'Informatique (effectué à Montpellier)
Co-encadrement : Pascal Neveu – INRA
Situation suivante : Thèse CIFRE Crédit Agricole, Brest

— **2014 (5 mois)** F. Philippe - Analyse de données de variations génétique dans les riz – Master2 bioinformatique Lumini (effectué à Montpellier)
co-encadrement : G. Sempere – Cirad
Situation suivante : Ingénieur Bioinformatique, INRA

— **2008 M2 Recherche (5 mois)** J. Wollbrett - Intégration automatique d'une ontologie de domaine dans un annuaire de service web bioinformatique : Biomoby – Master2 Bioinformatique UMII (effectué à Montpellier)
co-encadrement : M. Ruiz – Cirad
Situation suivante : Thèse CIRAD-Région LR

— **2007 M2 Recherche (5 mois)** S. Fromentin - Développement d'un Framework de services web pour l'interopérabilité de ressources agronomiques - Master2 Bioinformatique Orsay (effectué à Montpellier)
Situation suivante : Consultant Bioinformatique SS2I

— **2002 M2 Recherche (5 mois)** G. Droc – Développement d'un pipeline de traitement des séquences génomiques pour les mutants d'insertion chez le riz – Master2 Bioinformatique UMII (effectué à Montpellier)
situation suivante : Ingénieur Bioinformatique, Cirad

— **2001 M2 Recherche (6 mois** C. Tranchant – Développement d'un système d'information sur la traçabilité des échantillons OGM . DESS IAO UMII (effectué à Montpellier)
Situation suivante : Ingénieur Bioinformatique, IRD

Chapitre 8

Liste des publications

Je publie dans les thèmes (relatifs à l'intégration des données et des connaissances et dans le domaine de la bioinformatique. La plupart des publications sont indexées par :

— DBLP Computer Science Bibliography (20 entrées le 02/02/2020) :
`http://dblp.uni-trier.de/pers/hd/l/Larmande:Pierre`

— PUBMED (19 entrées le 02/02/2020) :
`https://www.ncbi.nlm.nih.gov/pubmed/?term=larmande+Pierre`

Mes publications sont listées ci-dessous par année de parution. Les impacts factor recensés dans cette liste sont issus des sites des journaux et mis à jour le 02/02/2020. Le rang des conférences est issu du site CORE (http ://103.1.187.206/core) et mis à jour le 02/02/2020. De nombreux articles ont été rédigés avec les doctorants et étudiants que je co-encadre. La règle pour l'ordre des noms est la suivante : le doctorant/étudiant en premier et les encadrants ou collaborateurs par ordre de taux de participation. Le dernier auteur est en général le responsable du projet. Dans le cas d'encadrement d'étudiants, il correspond au superviseur du travail. Les conférences internationales et nationales en biologie et bioinformatique ne produisent pas toujours des actes (proceedings), de même qu'elles ne sont pas classées. Par exemple la conférence Plant et Animal Genomes PAG rassemble plus de 3000 scientifiques depuis 25 ans sans produire de proceedings. C'est le cas également de JOBIM en France.

Thème	**Index de publication**	**Nombre de publication**
Journal	[1-24]	24
Conférence internationale	[25-37]	12
Conférence nationale	[38-41]	4
Workshop	[42-44]	3
Poster & démonstration	[45-57]	12
Ouvrage	[58]	1
Mémoire	[59]	1
Total		57

Publications avec comité de lecture

1. **Larmande P**, Do H, Wang Y. *OryzaGP : rice genes and proteins dataset for named-entity recognition.* Genomics Inform. 2019 ;17(2) :e17. Impact Factor : 1,5
2. Sempéré G, Pétel A, Rouard M, Frouin J, Hueber Y, De Bellis F, **Larmande P**. *Gigwa v2 – Extended and improved genotype investigator.* GigaScience, Volume 8, Issue 5, May 2019, giz051 Impact Factor : 7.31
3. Nti-Addae Y, Matthews D, Ulat V, Syed R, Sempere G, Petel A, Renner J, **Larmande P**, Guignon V, Jones E, Robbins K. *Benchmarking Database Systems for Genomic Selection Implementation.* 2019. Database. pii : baz096. Impact Factor : 3.978

4. Abbeloos R, Backlund JE, Basterrechea Salido M, Bauchet G, Benites-Alfaro O, Birkett C, Calaminos VC, Carceller P, Cornut G, Vasques Costa B, Edwards JD, Finkers R, Gao SY, Ghaffar M, Glaser P, Guignon V, Hok P, Kilian A, König P, Lagare JEB, Lange M, Laporte MA, **Larmande P**, LeBauer D, Lyon D, Marshall D, Matthews D, Milne I, Mistry N, Morales N, Mueller L, Neveu P, Papoutsoglou E, Pearce B, Perez-Masias I, Pommier C, Ramirez-Gonzalez RH, Rathore A, Raque AM, Raubach S, Rife T, Robbins K, Rouard M, Sarma C, Scholz U, Selby P, Sempéré G, Shaw P, Simon R, Soldevilla N, Stephen G, Sun Q, Tovar C, Uszynski G, Verouden M. *BrAPI - an Application Programming Interface for Plant Breeding Applications.* 2019. BioInformatics. pii :btz190. Impact Factor : 5.41
5. Juanillas V.M.J., Dereeper A., Beaume N., Droc G., Dizon J., Mendoza J.R., Perdon J.P., Mansueto L., Triplett L., Lang J., Zhou G., Ratharanjan K., Plale B., Haga J., Leach J.E., Ruiz M., Thomson M., Alexandrov N., **Larmande P.**, et al. *Rice Galaxy : an open resource for plant science.* Giga Science. 2019 1 ;8(5). pii : giz028. Impact Factor : 7.31
6. Venkatesan A., Tagny G., El Hassouni N., Chentli I., Guignon V., Jonquet C., Ruiz M., and **Larmande P.** *Agronomic Linked Data (AgroLD) : a Knowledge-based System to Enable Integrative Biology in Agronomy.* PLoS ONE 13(11) : e0198270. 2018. Impact Factor : 2.766
7. Cubry P., Tranchant-Dubreuil C., Thuillet A.C., Monat C., Ndjiondjop M.N., Labadie K., Cruaud C., Engelen S., Scarcelli N., Rhoné B., Burgarella C., Dupuy C., **Larmande P**., Wincker P., François O., Sabot F., and Vigouroux Y. *The Rise and Fall of African Rice Cultivation Revealed by Analysis of 246 New Genomes.* Curr Biol. Elsevier ; 2018 ;28 : 2274–2282.e6. Impact Factor : 9.201
8. Harper, Lisa ; Campbell, Jacqueline ; Cannon, Ethalinda K. S. ; Jung, Sook ; Poelchau, Monica ; Walls, Ramona ; Andorf, Carson ; Arnaud, Elizabeth ; Berardini, Tanya Z. ; Birkett, Clayton ; Cannon, Steve ; Carson, James ; Condon, Bradford ; Cooper, Laurel ; Dunn, Nathan ; Elsik, Christine G. ; Farmer, Andrew ; Ficklin, Stephen P. ; Grant, David ; Grau, Emily ; Herndon, Nic ; Hu, Zhi-Liang ; Humann, Jodi ; Jaiswal, Pankaj ; Jonquet, Clement ; Laporte, Marie-Angélique ; **Larmande, Pierre** ; Lazo, Gerard ; McCarthy, Fiona ; Menda, Naama ; Mungall, Christopher J. ; Munoz-Torres, Monica C. ; Naithani, Sushma ; Nelson, Rex ; Nesdill, Daureen ; Park, Carissa ; Reecy, James ; Reiser, Leonore ; Sanderson, Lacey-Anne ; Sen, Taner Z. ; Staton, Margaret ; Subramaniam, Sabarinath ; Tello-Ruiz, Marcela Karey ; Unda, Victor ; Unni, Deepak ; Wang, Liya ; Ware, Doreen ; Wegrzyn, Jill ; Williams, Jason ; Woodhouse, Margaret ; Yu, Jing ; Ware, Doreen. *AgBioData Consortium Recommendations for Sustainable Genomics and Genetics Databases for Agriculture.* Database. 2018 ; 1–7. Impact Factor : 3.978
9. Armin Scheben A., Chan K., Mansueto L., Mauleon R., **Larmande P.**, Alexandrov N., Wing R., McNally K., Quesneville H., Edwards D. *Progress in single access information systems for wheat and rice crop improvement.* Briefing in Bioinformatics. 2018 ; 4 :1-7 Impact Factor : 5.13
10. Jonquet C, Toulet A, Arnaud E, Aubin E, Dzalé-Yeumo E, Emonet V, Graybeal J, Laporte M-A, Musen M, Pesce V, **Larmande P**. *AgroPortal : an ontology repository for agronomy.* Comput. Electron. Agric. 2018 ; 144 :126–143 Impact Factor : 2.201
11. Dzale Yeumo, Esther ; Alaux, Michael ; Arnaud, Elizabeth ; Aubin, Sophie ; Baumann, Ute ; Buche, Patrice ; Cooper, Laurel ; Ćwiek-Kupczyńska, Hanna ; Davey, Robert P. ; Fulss, Richard Allan ; Jonquet, Clement ; Laporte, Marie-Angélique ; **Larmande, Pierre** ; Pommier, Cyril ; Protonotarios, Vassilis ; Reverte, Carmen ; Shrestha, Rosemary ; Subirats, Imma ; Venkatesan, Aravind ; Whan, Alex ; Quesneville, Hadi. *Developing data interoperability using standards : A wheat community use case.* F1000Research. 2017 ;6 :1843.
12. Cohen-Boulakia, Sarah ; Belhajjame, Khalid ; Collin, Olivier ; Chopard, Jérôme ; Froidevaux, Christine ; Gaignard, Alban ; Hinsen, Konrad ; **Larmande, Pierre** ; Bras, Yvan Le ; Lemoine,

Frédéric ; Mareuil, Fabien ; Ménager, Hervé ; Pradal, Christophe ; Blanchet, Christophe. *Scientific workflows for computational reproducibility in the life sciences : Status, challenges and opportunities.* Futur. Gener. Comput. Syst. 2017.75 : 284-298. Impact Factor : 2.786

13. The South Green Collaborators. *The South Green portal : a comprehensive resource for tropical and Mediterranean crop genomics.* Curr. Plant Biol. 2016. 7-8 : 6-9. Impact Factor : 1.68
14. Ngompé GT, Venkatesan A, Hassouni N, Ruiz M, **Larmande P**.*AgroLD API Une architecture orientée services pour l'extraction de connaissances dans la base de données liées AgroLD.* Lavoisier. 2016. 21 :133–58. Impact Factor : 1.046
15. Sempéré G, Philippe F, Dereeper A, Ruiz M, Sarah G, **Larmande P**. *Gigwa—Genotype investigator for genome- wide analyses.* Gigascience. 2016. 5 :25. Impact Factor : 7.31
16. Al-Tam, F., Adam, H., Dos Anjos, A., Lorieux, M., **Larmande, P.**, Ghesquière, A., Jouannic, S., and H-R Shahbazkia,*P-TRAP : a Panicle Traits Phenotyping Tool.* 2013, BMC Plant Biology, 13 :122-136. Impact Factor : 3.631
17. Wollbrett J, **Larmande P**, de Lamotte F, Ruiz M. *Clever generation of rich SPARQL queries from annotated relational schema : application to Semantic Web Service creation for biological databases.* BMC Bioinformatics. 2013. 14 :126-141.Impact Factor : 2.435
18. Lorieux, Mathias ; Blein, Mélisande ; Lozano, Jaime ; Bouniol, Mathieu ; Droc, Gaétan ; Diévart, Anne ; Périn, Christophe ; Mieulet, Delphine ; Lanau, Nadège ; Bès, Martine ; Rouvière, Claire ; Gay, Céline ; Piffanelli, Pietro ; **Larmande, Pierre** ; Michel, Corinne ; Barnola, Isabelle ; Biderre-Petit, Corinne ; Sallaud, Christophe ; Perez, Pascual ; Bourgis, Fabienne ; Ghesquière, Alain ; Gantet, Pascal ; Tohme, Joe ; Morel, Jean Benoit ; Guiderdoni, Emmanuel. *In-depth molecular and phenotypic characterization in a rice insertion line library facilitates gene identification through reverse and forward genetics approaches.* Plant Biotechnol. J. 2012 ;10 :555–568. Impact Factor : 7.443
19. Droc G, Périn C, Fromentin S, **Larmande P**. *OryGenesDB 2008 update : database interoperability for functional genomics of rice.* Nucleic Acids Res. 2009 ;37 :D992-D995. Impact factor : 9.202
20. **Larmande P**, Gay C, Lorieux M, Périn C, Bouniol M, Droc G, Sallaud C, Perez P, Barnola I, Biderre-Petit C, Martin J, Morel JB, Johnson AA, Bourgis F, Ghesquière, A, Ruiz M, Courtois B, Guiderdoni E. *Oryza Tag Line, a phenotypic mutant database for the Genoplante rice insertion line library.* Nucleic Acids Res. 2008 Jan ; 36(Database issue) :D1022-D1027. Impact factor : 9.202
21. Droc G, Ruiz M, **Larmande P**, Pereira A, Piffanelli P, Morel JB, et al. *OryGenesDB : a database for rice reverse genetics.* Nucleic Acids Res. 2006 ;34 :D736–D740. Impact factor : 9.202
22. Sallaud C., Gay C., **Larmande P.**, Bès M., Piffanelli P., Piégu B., Droc G., Regad F., Bourgeois E., Meynard D., Périn C., Sabau X., Ghesquière A., Delseny M., Glaszmann J.C., Guiderdoni, E. *High throughput T-DNA insertion mutagenesis in rice : A first step towards in silico reverse genetics.* Plant J. 2004 Aug ; 39(3) :450-64 Impact Factor : 5.468
23. Pugh T., Fouet O., Risterucci A.M., Brottier P., Abouladze M., Deletrez C., Courtois B., Clement D., **Larmande P.**, N'Goran J.A., Lanaud C., *A new cacao linkage map based on codominant markers : development and integration of 201 new microsatellite markers.* Theor Appl Genet. 2004. 108(6) :1151-61. 2004. Impact Factor ; 3.900
24. Sallaud C., Meynard D., van Boxtel J., Gay C., Bes M., Brizard J.P., **Larmande P.**, Ortega D., Raynal M., Portefaix M., Ouwerkerk P.B., Rueb S., Delseny M., Guiderdoni E., *Highly efficient production and characterization of T-DNA plants for rice (Oryza sativa L.) functional genomics.* Theor Appl Genet, 2003 ; 106 :1396-1408. Impact Factor ; 3.900

Communications internationales

25. Do Q and **Larmande P**. *Candidate gene prioritization using graph embedding*. IEEE-RIVF 2020. Ho Chi Minh, Vietnam. 6 pages.
26. **Larmande P**, Tagny Ngompé G, Ruiz M. *AgroLD : a Linked Data platform to understand plant genotype-phenotype interactions.* International Symposium on Integrative Bioinformatics 2019. Paris, France. 2 pages.
27. **Larmande P**. *The AgroLD project A Knowledge Graph-based Semantic Database for rice functional genomics.* International Symposium on Rice Functional Genomics ISRFG 2018. Tokyo, Japan 2 pages.
28. Mauleon R, McNally K, Mansueto L, Chebotarov D, Barboza L, Juanillas V, **Larmande P**, Droc G, Ruiz M, Dereeper A, Hamilton NR, Alexandrov N, Leung H, Wing R. *The international rice informatics consortium bioinformatics resources.* International Symposium on Rice Functional Genomics ISRFG 2018. Tokyo, Japan 2 pages.
29. Do H., Than K., and **Larmande P**.*Evaluating Named-Entity Recognition approaches in plant molecular biology.* Oral presentation at MIWAI 2018.Proceedings LNCS AI ; 11248. pp 219-225. 2018. Hanoi, Vietnam. 14 Pages
30. Do H., Than K., and **Larmande P**. *Comparative NER approaches in plant molecular biology.* CiCling 2018. 2018. Hanoi, Vietnam. 7 Pages. Rang B.
31. Pommier Cyril, Cornut Guillaume, Letellier Thomas, Michotey Célia, Neveu Pascal, Ruiz Manuel, **Larmande Pierre**, Kersey Paul J., Cwiek Hanna, Krajewski Pawel, Coppens Frederik, Finkers Richard, Laporte Marie-Angélique, Faria Daniel, Miguel Célia M., Chavez Ines, Adam-Blondon Anne-Françoise, Costa Bruno *Data standards for plant phenotyping : MIAPPE and its implementations.* PAG 2018. San Diego, USA. 2 pages.
32. **Larmande P.**, El Hassouni N. , Venkatesan A., Tagny G., Ruiz M.*The Agronomic Linked Data project (AgroLD) a knowledge network platform for rice.* International Symposium on Rice Functional Genomics ISRFG 2017. Sewon, Korea. 2 pages
33. Zevio S., El Hassouni N., Ruiz M. and **Larmande P**. *AgroLD indexing tools with ontological annotations.* Semantic Web for Life Science SWAT4LS 2016. Cambridge, UK. 4 pages
34. Jonquet C, Toulet A, Arnaud E, Aubin S, Yeumo ED, Emonet V, Graybeal J, Musen MA, Pommier C, **Larmande P**. *Reusing the NCBO BioPortal technology for agronomy to build AgroPortal.* Proceedings International Conference on Biomedical Ontology and BioCreative ICBO BioCreative 2016. CEUR Vol. 1747 Corvalis, USA. 6 pages.
35. Le Ngoc L, Tireau A, Venkatesan A, Neveu P, **Larmande P**.*Development of a knowledge system for Big Data : Case study to plant phenotyping data.* Proceedings. 6th Int. Conf. Web Intell. Min. Semant. WIMS 2016, Nimes, Fr. June 13-15, 2016. Nimes, France. ACM. p.27 :1-9.
36. **Larmande P**. *Ontology-based services and knowledge management in the Agronomic Domain.* 6th Research Data Alliance Conference RDA'2015. Paris, France. 2 pages.
37. Fromentin S., Droc G. and **Larmande P**. *A personalized, integrated system for rice functional genomics.* Network Tools and Applications in Biology NETTAB 2007, Pise, Italy. 4 pages.

Communications nationales

38. **Larmande P**, Tagny G, Ruiz M. *AgroLD : un graphe de connaissances pour la caractérisation des mécanismes moléculaires complexes impactant le phénome des plantes.* Conférence Francophone d'ingénierie des connaissances, IC 2019. Toulouse, France. 2 pages.
39. **Larmande P**.*Gigwa : Genotype Investigator for Genome Wide Analyses.* Journées ouvertes pour la Biologie, l'informatique et les Mathématiques. JOBIM 2018. Marseille, France. 2 pages

40. Venkatesan A, El Hassouni N, Philippe F, Pommier C, Quesneville H, Ruiz M and **Larmande P**. *Towards efficient data integration and knowledge management in the Agronomic domain.* Conference Francophone d'ingénierie des connaissances, IC 2015 Rennes, France. 6 pages.
41. Wollbrett J., **Larmande P**. and Ruiz M. *Intégration automatique d'une ontologie de domaine dans un annuaire Biomoby.* Journées ouvertes pour la Biologie, l'informatique et les Mathematiques. JOBIM 2009, Nantes, France. 8 pages.

Workshops

42. **Larmande P** *AgroLD : a Linked Data platform for agronomy.* International Symposium on Integrative Bioinformatics 2019. Paris, France. 2 pages. Workshop Open Data in Agrifood and Life Sciences : Models, Standards, Tools, Use Cases. Paris, France, 2019. 2 pages.
43. **Larmande P**. *Exposing French agronomic resources as Linked Open Data.* Présentatino orale au Workshop SoWeDO de la Conférence Francophone d'ingénierie des connaissances, IC 2016. Montpellier, France. 6 pages.
44. Wollbrett J, **Larmande P** and Ruiz M. *Towards Automatic Generation of Semantic Web services for relational Databases.* Présentation orale à l'International Workshop on Resources Discovery in conjunction with ESWC 2011. Heraclion, Greece. 6 pages.

Posters et démonstrations

45. **Larmande P**. *The AgroLD project update on a Knowledge Graph-based Semantic Database for rice functional genomics.* poster à International Symposium on Rice Functional Genomics ISRFG 2019. Taipei ,Taiwan 2 pages.
46. **Larmande P**.*Agronomic Linked Data (AgroLD) : a Knowledge-based System to Enable Integrative Biology in Agronomy.* Poster à JOBIM 2018. Marseille, France. 2 pages
47. Venkatesan A., Tagny G., El Hassouni N., Ruiz M., **Larmande P**. *The Agronomic Linked Data project.* Computer demo at Plant and Animal Genomes Conference PAG 2017. San Diego, USA. 2 pages
48. Sempere G., Phillippe F., Dereeper A., Ruiz M, Sarah G. and **Larmande P**. *Gigwa : Genotype Investigator for Genome Wide Analyses.* Computer demo at Plant and Animal Genomes Conference PAG 2017. San Diego, USA. 2 pages
49. Chentli I, **Larmande P**, Todorov K. *Construction d'un gold standard pour les données agronomiques.* Poster Conference Francophone d'Ingénierie des Connaissances, IC 2016. 251-254. Montpellier, France. 4 pages.
50. Robakowska Hyzorek D., Mirouze M., **Larmande P**. *Integration and Visualization of Epigenome and Mobilome Data in Crops.* Poster aux Journées ouvertes pour la Biologie, l'informatique et les Mathematiques JOBIM 2016. Lyon, France. 2 pages.
51. Le Ngoc L., Jouannic S. and **Larmande P**. *Développement d'un outil générique d'indexation pour optimiser l'exploitation de données biologiques.* Poster aux Journées ouvertes pour la Biologie, l'informatique et les Mathematiques JOBIM 2015. Clermont-Ferrant,France. 2 pages.
52. **Larmande P**. *Gigwa - Genotype Investigator for Genome Wide Analyses.* Computer demo at Plant and Animal Genomes Conference PAG 2015. San Diego, USA. 2 pages.
53. **Larmande P**, Venkatesan A, Jonquet C., Ruiz M. Sempere G., Valduriez P. *Enabling knowledge management in the Agronomic Domain .* Computer demo at Plant and Animal Genomes Conference PAG 2015. San Diego, USA. 2 pages.

54. **Larmande P**., Tranchant C., Libourel T., Mougenot I.*Intégration de données en génomique végétale.* Journées Ouvertes à la Biologie, l'Informatique et les Mathématiques, Satellite Workshop Ontologie, Grille et Intégration Sémantique pour la Biologie à la conference Biologie, l'informatique et les Mathematiques JOBIM 2007. Clermont-Ferrant (France) JOBIM 2005, Lyon, France. 8 pages.
55. **Larmande, P**. *Orylink : A Personalized Integrated System for Functional Genomic Analysis.* Computer demo at Plant and Animal Genomes Conference PAG 2009. San Diego, USA. 2 pages.
56. Maillol V, Bacilieri R, Sidibe Bocs S, Boursiquot J, Carrier G, Dereeper A, Droc G, Fleury C, **Larmande P**, Lecunff L, Péros JP, Pitollat B, Ruiz M, Sarah G, Sempéré G, Summo M, This P, and Dufayard JF. *Role of Galaxy in a bioinformatic plant breeding platform.* Poster at the Galaxy Community Conference 2012. Chicago, USA. 4 pages.
57. Fromentin S., Droc G. and **Larmande P**. *A personalized integrated system for rice functional genomic analysis.* Poster at the 5th International Symposium of Rice Functional Genomics ISRFG 2007. Tsukuba, Japon. 2 pages.

Édition d'ouvrages

58. **Larmande P.**, Mougenot I., Jonquet C., Libourel T., Ruiz M., Arnaud E. *Semantic Web for Biodiversity workshop proceedings*.Proceedings Semantics for Biodiversity Workshop.ESWC 2013. Montpellier, France. CEUR Vol 979 [1].

Mémoire

59. **Larmande P**. *Mutualiser et partager, un défi pour la génomique végétale.* Thèse de doctorat, Université Montpellier 2, Montpellier, France, 2007.

1. http://ceur-ws.org/Vol-979

Bibliographie

[1] Rafael ABBELOOS et al. « BrAPI - an Application Programming Interface for Plant Breeding Applications ». In : (mar. 2019). DOI : 10.1093/bioinformatics/btz190. (Visité le 28/03/2019).

[2] Manel ACHICHI, Zohra BELLAHSENE et Konstantin TODOROV. « Legato results for OAEI 2017 ». In : *Proceedings of the 12th International Workshop on Ontology Matching co-located with the 16th International Semantic Web Conference (ISWC 2017), Vienna, Austria, October 21, 2017.* 2017, p. 146-152. URL : http://ceur-ws.org/Vol-2032/oaei17_paper6.pdf.

[3] Manel ACHICHI et al. « Automatic Key Selection for Data Linking ». In : *20th International Conference on Knowledge Engineering and Knowledge Management - Volume 10024*. Springer-Verlag New York, Inc., 2016, p. 3-18.

[4] Manel ACHICHI et al. « Doing Web Data : from Dataset Recommendation to Data Linking ». In : *NoSQL Data Models*. willey. T. 1. Databases and Big Data SET. Olivier Pivert, juil. 2018, p. 57-91.

[5] Stein AERTS et al. « Gene prioritization through genomic data fusion ». In : *Nature Biotechnology* 24 (mai 2006), p. 537.

[6] A ALEXA et J RAHNENFUHRER. *topGO : Enrichment Analysis for Gene Ontology.* 2016. (Visité le 13/01/2017).

[7] Sophia ANANIADOU et al. « Event extraction for systems biology by text mining the literature ». eng. In : *Trends in Biotechnology* 28.7 (juil. 2010), p. 381-390. ISSN : 1879-3096. DOI : 10.1016/j.tibtech.2010.04.005.

[8] R ANICETO et R XAVIER. « Evaluating the Cassandra NoSQL Database Approach for Genomic Data Persistency ». In : *International ...* 2015 (2015). ISSN : 2314-436X. DOI : 10.1155/2015/502795.

[9] Erick ANTEZANA, Martin KUIPER et Vladimir MIRONOV. « Biological knowledge management : the emerging role of the Semantic Web technologies ». In : *Briefings in Bioinformatics* 10.4 (2009), p. 392 -407.

[10] Grigoris ANTONIOU et F HARMELEN. « Web ontology language : Owl ». In : *Handbook on ontologies* (2009).

[11] M. ASHBURNER et al. « Gene ontology : tool for the unification of biology. The Gene Ontology Consortium. » In : *Nat Genet* 25.1 (2000), p. 25-29. DOI : 10.1038/75556.

[12] Edward B. BARBIER et Jacob P. HOCHARD. « The Impacts of Climate Change on the Poor in Disadvantaged Regions ». en. In : *Review of Environmental Economics and Policy* 12.1 (fév. 2018). Publisher : Oxford Academic, p. 26-47. ISSN : 1750-6816. DOI : 10.1093/reep/rex023. URL : https://academic.oup.com/reep/article/12/1/26/4835833 (visité le 19/05/2020).

[13] Daniel BARRELL et al. « The GOA database in 2009 - An integrated Gene Ontology Annotation resource ». In : *Nucleic Acids Research* 37.SUPPL. 1 (2009). ISSN : 03051048. DOI : 10.1093/nar/gkn803.

[14] Marco BASALDELLA, Dario DE NART et Carlo TASSO. « Introducing Distiller : A Unifying Framework for Knowledge Extraction. » In : *IT@ LIA@ AI* IA* 1509 (2015).

[15] Marco BASALDELLA et al. « Entity recognition in the biomedical domain using a hybrid approach ». In : *Journal of Biomedical Semantics* 8.1 (2017), p. 1-14.

[16] François BELLEAU et al. « Bio2RDF : towards a mashup to build bioinformatics knowledge systems. » In : *Journal of biomedical informatics* 41.5 (2008), p. 706-16.

[17] Michael K. BERGMAN. « Sources and Classification of Semantic Heterogeneities ». English. In : *AI3 : : :Adaptive Information* (juin 2006). (Visité le 25/03/2019).

[18] Tim BERNERS-LEE et al. « The semantic web ». In : *Scientific american* 284.5 (2001), p. 29-37.

[19] Jiten BHAGAT et al. « BioCatalogue : a universal catalogue of web services for the life sciences ». In : *Nucleic Acids Research* 38.suppl$_2$ (mai 2010), W689-W694. ISSN : 0305-1048. DOI : 10.1093/nar/gkq394.

[20] Aaron BIRKLAND et Golan YONA. « BIOZON : a system for unification, management and analysis of heterogeneous biological data. » In : *BMC bioinformatics* 7 (jan. 2006), p. 70. ISSN : 1471-2105. DOI : 10.1186/1471-2105-7-70. (Visité le 10/04/2014).

[21] Christian BIZER. « D2R MAP – A Database to RDF Mapping Language ». In : *Proceedings of the 12th International World Wide Web Conference (WWW 2003)*. 2003.

[22] Christian BIZER et Andy SEABORNE. « D2RQ - treating non-RDF databases as virtual RDF graphs ». In : *the 3rd International Semantic Web Conference (ISWC 2004)*. 2004.

[23] J. A. BLAKE et al. « Gene ontology annotations and resources ». In : *Nucleic Acids Research* 41.D1 (2013). ISBN : 1362-4962 (Linking). ISSN : 03051048. DOI : 10.1093/nar/gks1050.

[24] Elizabeth I BOYLE et al. « GO : :TermFinder–open source software for accessing Gene Ontology information and finding significantly enriched Gene Ontology terms associated with a list of genes. » In : *Bioinformatics (Oxford, England)* 20.18 (déc. 2004), p. 3710-5. ISSN : 1367-4803. DOI : 10.1093/bioinformatics/bth456. (Visité le 09/03/2012).

[25] Richard BRUSKIEWICH et al. « Generation Challenge Programme (GCP) : standards for crop data ». In : *Omics : A Journal of Integrative Biology* 10.2 (2006), p. 215-219.

[26] Markus BUNDSCHUS et al. « Extraction of semantic biomedical relations from text using conditional random fields ». In : *BMC bioinformatics* 9 (avr. 2008), p. 207. ISSN : 1471-2105. DOI : 10.1186/1471-2105-9-207.

[27] William S BUSH, Scott M DUDEK et Marylyn D RITCHIE. « Biofilter : a knowledge-integration system for the multi-locus analysis of genome-wide association studies ». eng. In : *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* (2009), p. 368-379. ISSN : 2335-6928.

[28] Pier Luigi BUTTIGIEG et al. « The environment ontology in 2016 : bridging domains with increased scope, semantic density, and interoperation ». In : *Journal of Biomedical Semantics* 7.1 (2016). ISBN : 1332601600976. ISSN : 2041-1480. DOI : 10.1186/s13326-016-0097-6.

[29] Mary Elaine CALIFF et Raymond J MOONEY. « Relational learning of pattern-match rules for information extraction ». In : *Computational Linguistics* 4 (1999), p. 9-15.

[30] Alison CALLAHAN, José CRUZ-TOLEDO et Michel DUMONTIER. « Ontology-Based Querying with Bio2RDF's Linked Open Data. » In : *Journal of biomedical semantics* 4 Suppl 1.Suppl 1 (2013), S1.

[31] Emmanuel CASTANIER et al. « Semantic Annotation Workflow using Bio-Ontologies ». In : *Workshop on Crop Ontology and Phenotyping Data Interoperability*. 2014, p. 1.

[32] Wendy W. CHAPMAN et K. Bretonnel COHEN. « Current issues in biomedical text mining and natural language processing ». en. In : *Journal of Biomedical Informatics*. Biomedical Natural Language Processing 42.5 (oct. 2009), p. 757-759. ISSN : 1532-0464. DOI : 10.1016/j.jbi.2009.09.001. URL : http://www.sciencedirect.com/science/article/pii/S153204640900118X (visité le 17/04/2020).

[33] Dijun CHEN et al. « Bridging Genomics and Phenomics ». In : *Approaches in Integrative Bioinformatics - Towards the Virtual Cell*. 2014, p. 299-333. DOI : 10.1007/978-3-642-41281-3_11.

[34] Jing CHEN et al. « ToppGene Suite for gene list enrichment analysis and candidate gene prioritization ». eng. In : *Nucleic acids research* 37.Web Server issue (juil. 2009), W305-W311. ISSN : 1362-4962. DOI : 10.1093/nar/gkp427.

[35] Tao CHEN, Mingfen WU et Hexi LI. « A general approach for improving deep learning-based medical relation extraction using a pre-trained model and fine-tuning ». In : *Database : The Journal of Biological Databases and Curation* 2019 (déc. 2019). ISSN : 1758-0463. DOI : 10.1093/database/baz116. URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6892305/.

[36] Marcus C CHIBUCOS et al. « Standardized description of scientific evidence using the Evidence Ontology (ECO) ». In : *Database : the journal of biological databases and curation* 2014 (juil. 2014). Publisher : Oxford University Press, bau075. ISSN : 1758-0463. DOI : 10.1093/database/bau075. URL : https://pubmed.ncbi.nlm.nih.gov/25052702.

[37] F. CIRAVEGNA et al. « LearningPinocchio : adaptive information extraction for real world applications ». In : *Natural Language Engineering* 10 (1999), p. 145-165.

[38] Michael COCHEZ et al. « Global RDF Vector Space Embeddings. » In : *The Semantic Web - ISWC 2017 - 16th International Semantic Web Conference, Vienna, Austria, October 21-25, 2017, Proceedings, Part I*. 2017, p. 190-207. DOI : 10.1007/978-3-319-68288-4_12.

[39] Sarah COHEN-BOULAKIA et Ulf LESER. « Next Generation Data Integration for the Life Sciences ». In : *ICDE* (2010).

[40] Remi COLETTA et al. « Public Data Integration with WebSmatch ». In : *Proceedings of the First International Workshop on Open Data*. WOD '12. ACM, 2012, p. 5-12.

[41] Bradford CONDON et al. « Tripal Developer Toolkit ». eng. In : *Database : The Journal of Biological Databases and Curation* 2018 (2018). ISSN : 1758-0463. DOI : 10.1093/database/bay099.

[42] Laurel COOPER et al. « The Planteome database : An integrated resource for reference ontologies, plant genomics and phenomics ». In : *Nucleic Acids Research* 46.D1 (2018). ISSN : 13624962. DOI : 10.1093/nar/gkx1152.

[43] P. CORBETT et J. BOYLE. « Improving the learning of chemical-protein interactions from literature using transfer learning and specialized word embeddings ». eng. In : *Database : The Journal of Biological Databases and Curation* 2018 (2018). ISSN : 1758-0463. DOI : 10.1093/database/bay066.

[44] Olivier CORBY, Rose DIENG-KUNTZ et Catherine FARON-ZUCKER. « Querying the Semantic Web with Corese Search Engine. » In : *Proceedings of the 16th Eureopean Conference on Artificial Intelligence, ECAI'2004, including Prestigious Applicants of Intelligent Systems, PAIS 2004, Valencia, Spain, August 22-27, 2004*. 2004, p. 705-709.

[45] Richard G CÔTÉ et al. « The Ontology Lookup Service, a lightweight cross-platform tool for controlled vocabulary queries. » In : *BMC bioinformatics* 7 (2006), p. 97.

[46] Nadine CULLOT, Raji GHAWI et Kokou YÉTONGNON. « DB2OWL : A Tool for Automatic Database-to-Ontology Mapping 2 Database to Ontology Mappings : DB2OWL Tool ». In : (2007), p. 1-4.

[47] Hanjun DAI et al. « Sequence2Vec : a novel embedding approach for modeling transcription factor binding affinity landscape ». eng. In : *Bioinformatics (Oxford, England)* 33.22 (nov. 2017), p. 3575-3583. ISSN : 1367-4811. DOI : 10.1093/bioinformatics/btx480.

[48] Petr DANECEK et al. « The variant call format and VCFtools. » In : *Bioinformatics (Oxford, England)* 27.15 (2011), p. 2156-8. ISSN : 1367-4811. DOI : 10.1093/bioinformatics/btr330.

[49] Susan B. DAVIDSON et al. « K2Kleisli and GUS : Experiments in Integrated Access to Genomic Data Sources ». In : *IBM Systems Journal* 40.2 (2001), p. 512-31.

[50] Rahul C DEO et al. « Prioritizing causal disease genes using unbiased genomic features ». eng. In : *Genome biology* 15.12 (déc. 2014), p. 534-534. ISSN : 1474-760X. DOI : 10.1186/s13059-014-0534-8.

[51] Alexis DEREEPER et al. « SNiPlay3 : a web-based application for exploration and large scale analyses of genomic variations. » In : *Nucleic acids research* 43.W1 (2015), W295-300.

[52] Alexis DEREEPER et al. « SNiPlay3 : a web-based application for exploration and large scale analyses of genomic variations. » In : *Nucleic acids research* 43.W1 (2015), W295-300. ISSN : 1362-4962. DOI : 10.1093/nar/gkv351.

[53] Jacob DEVLIN et al. « Bert : Pre-training of deep bidirectional transformers for language understanding ». In : *arXiv preprint arXiv :1810.04805* (2018).

[54] Anastasia DIMOU et al. « RML : A Generic Language for Integrated RDF Mappings of Heterogeneous Data ». In : *Proceedings of the Workshop on Linked Data on the Web co-located with the 23rd International World Wide Web Conference (WWW 2014), Seoul, Korea, April 8, 2014*. Sous la dir. de Christian BIZER et al. T. 1184. CEUR Workshop Proceedings. CEUR-WS.org, 2014. URL : http://ceur-ws.org/Vol-1184/ldow2014_paper_01.pdf.

[55] Marisa P. DOLLED-FILHART et al. « Computational and Bioinformatics Frameworks for Next-Generation Whole Exome and Genome Sequencing ». In : *The Scientific World Journal* 2013 (jan. 2013), p. 730210. ISSN : 2356-6140. DOI : 10.1155/2013/730210. URL : https://doi.org/10.1155/2013/730210.

[56] G DROC et al. « OryGenesDB 2008 update : database interoperability for functional genomics of rice ». In : *Nucleic Acids Research* 37.Database issue (2009), p. D992-D995. ISSN : 1362-4962. DOI : 10.1093/nar/gkn821.

[57] Gaëtan DROC et al. « OryGenesDB 2008 update : database interoperability for functional genomics of rice ». In : *Nucleic Acids Research* 37.Database issue (2009), p. D992-5.

[58] Esther DZALE YEUMO et al. « Developing data interoperability using standards : A wheat community use case ». In : *F1000Research* 6 (2017), p. 1843.

[59] Ramez ELMASRI et Shamkant B. NAVATHE. *Fundamentals of Database Systems (5th Edition)*. Boston, MA, USA : Addison-Wesley Longman Publishing Co., Inc., 2006. ISBN : 0-321-36957-2.

[60] Sarah ELSHAL et al. « Beegle : From literature mining to disease-gene discovery ». In : *Nucleic Acids Research* 44.2 (2016), e18.

[61] T. ETZOLD, A. ULYANOV et P. ARGOS. « SRS : information retrieval system for molecular biology data banks. » In : *Methods Enzymol* 266 (1996), p. 114-28.

[62] Daniel FARIA et al. « The AgreementMakerLight ontology matching system ». In : *Lecture Notes in Computer Science* 8185 LNCS (2013), p. 527-541.

[63] Farzad FARNOUD, Minji KIM et Olgica MILENKOVIC. « HyDRA : gene prioritization via hybrid distance-score rank aggregation ». In : *Bioinformatics* 31.7 (nov. 2014), p. 1034-1043. ISSN : 1367-4803. DOI : 10.1093/bioinformatics/btu766. (Visité le 06/04/2019).

[64] Alfio FERRARA, Andriy NIKOLOV et François SCHARFFE. « Data Linking for the Semantic Web ». In : *International Journal on Semantic Web and Information Systems (IJSWIS)* 7.3 (2011), p. 46-76.

[65] S. P. FICKLIN et al. « Tripal : a construction toolkit for online genome databases ». In : *Database* 2011.0 (sept. 2011), bar044-bar044. ISSN : 1758-0463. DOI : 10.1093/database/bar044. (Visité le 21/05/2018).

[66] Kristofer FRANZÉN et al. « Protein names and how to find them. » In : *International journal of medical informatics* 67.1-3 (2002), p. 49-61.

[67] Matthew L FREEDMAN et al. « Principles for the post-GWAS functional characterization of cancer risk loci ». In : *Nature Genetics* 43 (mai 2011), p. 513.

[68] Robert T. FURBANK et Mark TESTER. « Phenomics – technologies to relieve the phenotyping bottleneck ». In : *Trends in Plant Science* 16.12 (2011), p. 635 -644. ISSN : 1360-1385. DOI : https://doi.org/10.1016/j.tplants.2011.09.005.

[69] Matteo GABETTA et al. « BigQ : a NoSQL based framework to handle genomic variants in i2b2 ». In : *BMC Bioinformatics* 16 (2015). Publisher : BMC Bioinformatics, p. 415. ISSN : 1471-2105. DOI : 10.1186/s12859-015-0861-0.

[70] Santhosh Kumar GAJENDRAN. « A Survey on NoSQL Databases ». In : *University of Illinois* (2012).

[71] Fabien GANDON, Catherine FARON-ZUCKER et Olivier CORBY. *Le Web sémantique : comment lier les données et les schémas sur le web ?* Dunod, 2012, p. 1-81,88-125,163-166.

[72] Yael GARTEN, Adrien COULET et Russ B ALTMAN. « Recent progress in automatically extracting information from the pharmacogenomic literature ». In : *Pharmacogenomics* 11.10 (oct. 2010). Publisher : Future Medicine, p. 1467-1489. ISSN : 1462-2416. DOI : 10.2217/pgs.10.136. URL : https://www.futuremedicine.com/doi/10.2217/pgs.10.136 (visité le 17/04/2020).

[73] Martin GERNER, Goran NENADIC et Casey M BERGMAN. « LINNAEUS : A species name identification system for biomedical literature ». In : *BMC Bioinformatics* 11.1 (2010), p. 85.

[74] Belinda GIARDINE et al. « Galaxy : A platform for interactive large-scale genome analysis ». In : *Genome Research* 15.10 (2005), p. 1451-1455.

[75] John M. GIORGI et Gary D. BADER. « Transfer learning for biomedical named entity recognition with neural networks ». eng. In : *Bioinformatics (Oxford, England)* 34.23 (2018), p. 4087-4094. ISSN : 1367-4811. DOI : 10.1093/bioinformatics/bty449.

[76] Jeremy GOECKS, Anton NEKRUTENKO et James TAYLOR. « Galaxy : a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences ». In : *Genome Biology* 11.8 (2010), p. 1-13.

[77] Christine GOLBREICH et al. « OBO and OWL : Leveraging Semantic Web Technologies for the Life Sciences ». In : *ISWC 2007*. 2007, p. 169-182.

[78] Katarina GROLINGER et al. « Data management in cloud environments : NoSQL and NewSQL data stores ». In : *Journal of Cloud Computing : Advances, Systems and Applications* 2.1 (2013), p. 22. ISSN : 2192-113X. DOI : 10.1186/2192-113X-2-22.

[79] L. M. HAAS et al. « DiscoveryLink : a system for integrated access to life sciences data sources ». In : *IBM Syst. J.* 40.2 (2001), p. 489-511. ISSN : 0018-8670.

[80] Maryam HABIBI et al. « Deep learning with word embeddings improves biomedical named entity recognition ». In : *Bioinformatics* 33.14 (2017), p. i37-i48.

[81] Alon Y. HALEVY. « Answering queries using views : A survey ». In : *The VLDB Journal* (2001). (Visité le 04/02/2013).

[82] Chantal HAMELIN et al. « TropGeneDB, the multi-tropical crop information system updated and extended ». In : *Nucleic acids research* (2012), gks1105. DOI : 10.1093/nar/gks1105.

[83] Ian HARROW et al. « Matching disease and phenotype ontologies in the ontology alignment evaluation initiative ». In : *Journal of Biomedical Semantics* 8.1 (2017), p. 1-13. ISSN : 20411480. DOI : 10.1186/s13326-017-0162-9.

[84] Keywan HASSANI-PAK et al. « Developing integrated crop knowledge networks to advance candidate gene discovery ». In : *Applied & Translational Genomics* 11 (2016), p. 18-26.

[85] Christian Theil HAVE et Lars Juhl JENSEN. « Databases and ontologies Are graph databases ready for bioinformatics ? » In : *Bioinformatics* 29.24 (2013), p. 3107-3108.

[86] Philipp HEIM et al. « 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)*. T. 5887 LNCS. Citation Key : Heim2009 ISSN : 03029743. 2009, p. 182-187. ISBN : 3-642-10542-4. DOI : 10.1007/978-3-642-10543-2_21.

[87] Kristina M. HETTNE et al. « A dictionary to identify small molecules and drugs in free text ». In : *Bioinformatics* 25.22 (2009), p. 2983-2991.

[88] L. HIRSCHMAN et R. GAIZAUSKAS. « Natural Language Question Answering : The View from Here ». In : *Nat. Lang. Eng.* 7.4 (2001), p. 275-300.

[89] S. K. HONG et Jae-Gil LEE. « DTranNER : biomedical named entity recognition with deep learning-based label-label transition model ». In : *BMC Bioinformatics* 21.1 (fév. 2020), p. 53. ISSN : 1471-2105. DOI : 10.1186/s12859-020-3393-1. URL : https://doi.org/10.1186/s12859-020-3393-1 (visité le 17/04/2020).

[90] Lin HOU et Hongyu ZHAO. « A review of post-GWAS prioritization approaches ». In : *Frontiers in Genetics* 4.DEC (2013). ISBN : 1664-8021 (Print)\r1664-8021 (Linking), p. 2009-2014. ISSN : 16648021. DOI : 10.3389/fgene.2013.00280.

[91] David HOULE, Diddahally R. GOVINDARAJU et Stig OMHOLT. « Phenomics : the next challenge ». In : *Nature Reviews Genetics* 11 (nov. 2010), p. 855.

[92] Kyu-Baek HWANG et al. « Comparative analysis of whole-genome sequencing pipelines to minimize false negative findings ». In : *Scientific Reports* 9.1 (mar. 2019), p. 3219. ISSN : 2045-2322. DOI : 10.1038/s41598-019-39108-2. URL : https://doi.org/10.1038/s41598-019-39108-2.

[93] Hyunchul JANG et al. « Finding the evidence for protein-protein interactions from PubMed abstracts ». In : *Bioinformatics* 22.14 (2006), e220-e226.

[94] Lars Juhl JENSEN, Jasmin SARIC et Peer BORK. « Literature mining for the biologist : from information retrieval to biological discovery ». In : *Nature Reviews Genetics* 7.2 (fév. 2006), p. 119-129. ISSN : 1471-0064. DOI : 10.1038/nrg1768. URL : https://doi.org/10.1038/nrg1768.

[95] Anja JENTZSCH et al. « Silk – Generating RDF Links while publishing or consuming Linked Data ». In : *Proceedings of ISWC* (2010).

[96] Clément JONQUET et al. « AgroPortal : A vocabulary and ontology repository for agronomy ». In : *Computers and Electronics in Agriculture* 144.October 2016 (2018), p. 126-143.

[97] Clément JONQUET et al. « Indexation et intégration de ressources textuelles à l ' aide d ' ontologies : application au domaine biomédical ». In : *IC2010* (2010), p. 1-12.

[98] Clement JONQUET et al. « Reusing the NCBO BioPortal technology for agronomy to build AgroPortal ». In : *7th International Conference on Biomedical Ontologies* 1747 (2016).

[99] Jelena JOVANOVIĆ et Ebrahim BAGHERI. « Semantic annotation in biomedicine : the current landscape ». In : *Journal of Biomedical Semantics* 8.1 (2017), p. 44.

[100] Simon JUPP et al. « The EBI RDF platform : linked open data for the life sciences. » In : *Bioinformatics (Oxford, England)* (2014), p. 1-2.

[101] Tim KACPROWSKI, Nadezhda T DONCHEVA et Mario ALBRECHT. « NetworkPrioritizer : a versatile tool for network-based prioritization of candidate disease genes or other molecules ». eng. In : *Bioinformatics (Oxford, England)* 29.11 (juin 2013), p. 1471-1473. ISSN : 1367-4811. DOI : 10.1093/bioinformatics/btt164.

[102] Ning KANG et al. « Knowledge-based extraction of adverse drug events from biomedical text ». In : *BMC Bioinformatics* 15.1 (mar. 2014), p. 64. ISSN : 1471-2105. DOI : 10.1186/1471-2105-15-64. URL : https://doi.org/10.1186/1471-2105-15-64 (visité le 20/04/2020).

[103] Peter D KARP, Thomas J LEE et Valerie WAGNER. « BioWarehouse : Relational Integration of Eleven Bioinformatics Databases and Formats ». In : (2008), p. 5-7.

[104] Ji-Sung KIM, Xin GAO et Andrey RZHETSKY. « RIDDLE : Race and ethnicity Imputation from Disease history with Deep LEarning ». eng. In : *PLoS computational biology* 14.4 (avr. 2018), e1006106-e1006106. ISSN : 1553-7358. DOI : 10.1371/journal.pcbi.1006106.

[105] Jin-Dong KIM et Kevin Bretonnel COHEN. « Natural language query processing for SPARQL generation : A prototype system for SNOMED-CT ». In : *Proceedings of the BioLINK SIG*. 2013, p. 32-38.

[106] Paweł KRAJEWSKI et al. « Towards recommendations for metadata and data handling in plant phenotyping ». In : *Journal of Experimental Botany* 66.18 (2015), p. 5417-5427.

[107] Martin KRALLINGER, Florian LEITNER et Obdulia RABAL. « Overview of the chemical compound and drug name recognition (CHEMDNER) task ». In : *Proceedings of the Fourth BioCreative Challenge Evaluation Workshop* 2 (2013), p. 2-33.

[108] Maxat KULMANOV et al. « Vec2SPARQL : integrating SPARQL queries and knowledge graph embeddings ». In : *bioRxiv* (jan. 2018), p. 463778. DOI : 10.1101/463778.

[109] Ajay Anand KUMAR et al. « pBRIT : gene prioritization by correlating functional and phenotypic annotations through integrative data fusion ». eng. In : *Bioinformatics (Oxford, England)* 34.13 (juil. 2018), p. 2254-2262. ISSN : 1367-4811. DOI : 10.1093/bioinformatics/bty079.

[110] Jacob KÖHLER et al. « Graph-based analysis and visualization of experimental results with ONDEX ». en. In : *Bioinformatics* 22.11 (juin 2006), p. 1383-1390. ISSN : 1367-4803. DOI : 10.1093/bioinformatics/btl081. (Visité le 26/03/2019).

[111] John LAFFERTY, Andrew MCCALLUM et Fernando C N PEREIRA. « Conditional random fields : Probabilistic models for segmenting and labeling sequence data ». In : *ICML '01 Proceedings of the Eighteenth International Conference on Machine Learning* 8.June (2001), p. 282-289.

[112] Camille LAIBE et al. « Identifiers. org : integration tool for heterogeneous datasets ». In : *Dils 2014* (2014). Citation Key : Laibe2014, p. 14. DOI : 10.6084/m9.figshare.1232122.v1.

[113] Guillaume LAMPLE et al. « Neural Architectures for Named Entity Recognition ». In : (2016).

[114] P LARMANDE et al. « Oryza Tag Line, a phenotypic mutant database for the Genoplante rice insertion line library ». In : *Nucleic Acids Research* 36.Database issue (2008), p. D1022-1027. ISSN : 1362-4962. DOI : 10.1093/nar/gkm762.

[115] Pierre LARMANDE, Huy DO et Yue WANG. « OryzaGP : rice gene and protein dataset for named-entity recognition ». In : *Genomics & Informatics* 17.2 (juin 2019), e17. ISSN : 1598-866X. DOI : 10.5808/GI.2019.17.2.e17.

[116] Pierre LARMANDE et Kazim Muhammed JIBRIL. « Enabling Fast Annotation Process With Table2Annotation Tool ». In : *bioRxiv* (avr. 2020), p. 2020.04.03.023069. DOI : 10.1101/2020.04.03.023069. URL : https://www.biorxiv.org/content/10.1101/2020.04.03.023069v1 (visité le 06/05/2020).

[117] Duc Hau LE et Yung Keun KWON. « GPEC : A Cytoscape plug-in for random walk-based gene prioritization and biomedical evidence collection ». In : *Computational Biology and Chemistry* 37 (2012), p. 17-23.

[118] Duc Hau LE et Yung Keun KWON. « Neighbor-favoring weight reinforcement to improve random walk-based disease gene prioritization ». In : *Computational Biology and Chemistry* 44 (2013), p. 1-8.

[119] ROBERT LEAMAN et GRACIELA GONZALEZ. « Banner : an Executable Survey of Advances in Biomedical Named Entity Recognition ». In : *Biocomputing 2008* 663 (2007), p. 652-663.

[120] Insuk LEE et al. « Prioritizing candidate disease genes by network-based boosting of genome-wide association data ». eng. In : *Genome research* 21.7 (juil. 2011), p. 1109-1121. ISSN : 1549-5469. DOI : 10.1101/gr.118992.110.

[121] Jinhyuk LEE et al. « BioBERT : a pre-trained biomedical language representation model for biomedical text mining ». In : *Bioinformatics* 36.4 (sept. 2019), p. 1234-1240. ISSN : 1367-4803. DOI : 10.1093/bioinformatics/btz682. URL : https://doi.org/10.1093/bioinformatics/btz682 (visité le 04/02/2020).

[122] Maurizio LENZERINI. « Data Integration : A Theoretical Perspective ». In : *Proceedings of the Twenty-first ACM SIGACT-SIGMOD-SIGART Symposium on Principles of Database Systems, June 3-5, Madison, Wisconsin, USA*. 2002, p. 233-246. DOI : 10.1145/543613.543644.

[123] Sabina LEONELLI et al. « Data management and best practice for plant science ». In : *Nature Publishing Group* 3.June (2017), p. 1-4.

[124] « Data Integration in the Life Sciences, Third International Workshop, DILS 2006, Hinxton, UK, July 20-22, 2006, Proceedings ». In : Lecture Notes in Computer Science 4075 (2006). Sous la dir. d'Ulf LESER, Felix NAUMANN et Barbara A. ECKMAN. DOI : 10.1007/11799511. URL : https://doi.org/10.1007/11799511.

[125] Yongjin LI et Jinyan LI. « Disease gene identification by random walk on multigraphs merging heterogeneous genomic and phenotype data ». eng. In : *BMC genomics* 13 Suppl 7.Suppl 7 (déc. 2012), S27-S27. ISSN : 1471-2164. DOI : 10.1186/1471-2164-13-S7-S27.

[126] Yongjin LI et Jagdish C PATRA. « Integration of multiple data sources to prioritize candidate genes using discounted rating system ». eng. In : *BMC bioinformatics* 11 Suppl 1.Suppl 1 (jan. 2010), S20-S20. ISSN : 1471-2105. DOI : 10.1186/1471-2105-11-S1-S20.

[127] Yu LI et al. « DEEPre : sequence-based enzyme EC number prediction by deep learning ». eng. In : *Bioinformatics (Oxford, England)* 34.5 (mar. 2018), p. 760-769. ISSN : 1367-4811. DOI : 10.1093/bioinformatics/btx680.

[128] Sangrak LIM et Jaewoo KANG. « Chemical-gene relation extraction using recursive neural network ». eng. In : *Database : The Journal of Biological Databases and Curation* 2018 (2018). ISSN : 1758-0463. DOI : 10.1093/database/bay060.

[129] Haibin LIU, Ravikumar KOMANDUR et Karin VERSPOOR. « From graphs to events : A subgraph matching approach for information extraction from biomedical text ». In : *Proceedings of the BioNLP Shared Task 2011 Workshop*. Association for Computational Linguistics. 2011, p. 164-172.

[130] Sijia LIU et al. « Extracting chemical-protein relations using attention-based neural networks ». eng. In : *Database : The Journal of Biological Databases and Curation* 2018 (2018). ISSN : 1758-0463. DOI : 10.1093/database/bay102.

[131] Vanessa LOPEZ et al. « Is Question Answering Fit for the Semantic Web ? : A Survey ». In : *Semant. web* 2.2 (2011), p. 125-155.

[132] L E Ngoc LUYEN et al. « Development of a Knowledge System for Big Data : Case Study to Plant Phenotyping Data ». In : *Proceedings of the 6th International Conference on Web Intelligence, Mining and Semantics*. T. 27. WIMS '16. ACM, 2016, p. 1-9.

[133] Artem LYSENKO et al. « Representing and querying disease networks using graph databases ». In : *BioData Mining* 9.1 (2016), p. 23.

[134] Michele MAGRANE et Uni Prot CONSORTIUM. « UniProt Knowledgebase : A hub of integrated protein data ». In : *Database* 2011 (2011). ISSN : 17580463. DOI : 10.1093/database/bar009.

[135] Ioana MANOLESCU et al. « Efficient Querying of Distributed Resources in Mediator Systems ». In : *CoopIS/DOA/ODBASE* (2002), p. 468-485.

[136] Teri A. MANOLIO et al. « Finding the missing heritability of complex diseases ». In : *Nature* 461 (oct. 2009), p. 747.

[137] Ganiraju MANYAM et al. « Relax with CouchDB - Into the non-relational DBMS era of bioinformatics. » In : *Genomics* 100.1 (mai 2012). Publisher : Elsevier Inc., p. 1-7. ISSN : 1089-8646. DOI : 10.1016/j.ygeno.2012.05.006. (Visité le 05/06/2012).

[138] Sérgio MATOS et Rui ANTUNES. « Protein-Protein Interaction Article Classification Using a Convolutional Recurrent Neural Network with Pre-trained Word Embeddings ». In : *Journal of Integrative Bioinformatics* 14.4 (déc. 2017). ISSN : 1613-4516. DOI : 10.1515/jib-2017-0055. URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6042813/ (visité le 20/04/2020).

[139] Takashi MATSUMOTO et al. « The Nipponbare genome and the next-generation of rice genomics research in Japan ». In : *Rice* 9.1 (juil. 2016), p. 33. ISSN : 1939-8433. DOI : 10.1186/s12284-016-0107-4. (Visité le 07/03/2019).

[140] Soumia MELZI et Clement JONQUET. « Scoring semantic annotations returned by the NCBO Annotator ». In : *7th International Semantic Web Applications and Tools for Life Sciences, SWAT4LS'14*. CEUR Workshop Proceedings., 2014, Vol. 1320 pp. 15.

[141] Franck MICHEL, Catherine FARON ZUCKER et Johan MONTAGNAT. « Bridging the Semantic Web and NoSQL Worlds : Generic SPARQL Query Translation and Application to MongoDB ». In : *Transactions on Large-Scale Data- and Knowledge-Centered Systems XL*. T. 11360. LNCS. Springer, jan. 2019, p. 125-165. URL : https://hal.archives-ouvertes.fr/hal-01926379.

[142] Franck MICHEL, Johan MONTAGNAT et Catherine FARON-ZUCKER. « A survey of RDB to RDF translation approaches and tools ». In : *HAL* May (2014).

[143] Franck MICHEL et al. « xR2RML : Non-Relational Databases to RDF Mapping Language ». In : *HAL* (2015).

[144] Tomáš MIKOLOV, Wen-tau YIH et Geoffrey ZWEIG. « Linguistic regularities in continuous space word representations ». In : *Proceedings of the 2013 conference of the north american chapter of the association for computational linguistics : Human language technologies*. 2013, p. 746-751.

[145] Tomas MIKOLOV et al. « word2vec ». In : *URL https ://code. google. com/p/word2vec* (2013).

[146] Iain MILNE et al. « Flapjack–graphical genotype visualization. » In : *Bioinformatics (Oxford, England)* 26.24 (2010), p. 3133-4.

[147] Dan MOLDOVAN et al. « Performance Issues and Error Analysis in an Open-Domain Question Answering System ». In : *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*. 2002, p. 33-40.

[148] Cécile MONAT et al. « TOGGLE : toolbox for generic NGS analyses. » In : *BMC bioinformatics* 16 (2015), p. 374.

[149] A B M MONIRUZZAMAN et Syed Akhter HOSSAIN. « NoSQL Database : New Era of Databases for Big data Analytics - Classification, Characteristics and Comparison ». In : *CoRR* abs/1307.0.4 (2013). arXiv : 1307.0191 ISBN : 2005-4270, p. 1-14. ISSN : 2005-4270.

[150] Fantine MORDELET et Jean-Philippe VERT. « ProDiGe : Prioritization Of Disease Genes with multitask machine learning from positive and unlabeled examples ». eng. In : *BMC bioinformatics* 12 (oct. 2011), p. 389-389. ISSN : 1471-2105. DOI : 10.1186/1471-2105-12-389.

[151] Yves MOREAU et Léon Charles TRANCHEVENT. « Computational tools for prioritizing candidate genes : Boosting disease gene discovery ». In : *Nature Reviews Genetics* 13.8 (2012), p. 523-536.

[152] Nagarajan NATARAJAN et Inderjit S DHILLON. « Inductive matrix completion for predicting gene-disease associations ». eng. In : *Bioinformatics (Oxford, England)* 30.12 (juin 2014), p. i60-i68. ISSN : 1367-4811. DOI : 10.1093/bioinformatics/btu269.

[153] Mariana NEVES et Ulf LESER. « Question answering for Biology ». In : *Methods* 74 (2015), p. 36-46.

[154] Pascal NEVEU et al. « Dealing with multi-source and multi-scale information in plant phenomics : the ontology-driven Phenotyping Hybrid Information System ». In : *New Phytologist* 0.0 (). DOI : 10.1111/nph.15385.

[155] a.C.N. NGOMO et Sörer AUER. « Limes-a time-efficient approach for large-scale link discovery on the web of data ». In : *Proceedings of IJCAI* (2011), p. 2312-2317.

[156] Andriy NIKOLOV et al. « Combining RDF Graph Data and Embedding Models for an Augmented Knowledge Graph ». In : *Companion of the The Web Conference 2018 on The Web Conference 2018*. International World Wide Web Conferences Steering Committee. 2018, p. 977-980.

[157] Henrik NORDBERG et al. « BioPig : A Hadoop-based analytic toolkit for large-scale sequence data ». In : *Bioinformatics* 29.23 (2013). ISBN : 1367-4811 (Electronic), p. 3014-3019. ISSN : 13674803. DOI : 10.1093/bioinformatics/btt528.

[158] Natalya F NOY et al. « BioPortal : ontologies and integrated data resources at the click of a mouse ». In : *Nucleic Acids Research* 37.Web Server issue (2009), W170-173.

[159] Yaw NTI-ADDAE et al. « Benchmarking Database Systems for Genomic Selection Implementation ». In : *bioRxiv* (jan. 2019), p. 519017. DOI : 10.1101/519017.

[160] Yaw NTI-ADDAE et al. « Benchmarking database systems for Genomic Selection implementation ». In : *Database* 2019.baz096 (sept. 2019). ISSN : 1758-0463. DOI : 10.1093/database/baz096. URL : https://doi.org/10.1093/database/baz096 (visité le 05/04/2020).

[161] Edison ONG et al. « Ontobee : A linked ontology data server to support ontology term dereferencing, linkage, query and integration ». In : *Nucleic Acids Research* (2016), gkw918.

[162] James M OSTELL. « Entrez : The NCBI Search and Discovery Engine ». In : (2012), p. 1-4.

[163] Lorena OTERO-CERDEIRA, Francisco J. RODRÍGUEZ-MARTÍNEZ et Alma GÓMEZ-RODRÍGUEZ. « Ontology matching : A literature review ». In : *Expert Systems with Applications* 42.2 (2015), p. 949-971.

[164] Pablo PAREJA-TOBES et al. « Bio4j : a high-performance cloud-enabled graph-based data platform ». In : *BioXriv* (2015), p. 1-11.

[165] Jeffrey PENNINGTON, Richard SOCHER et Christopher D MANNING. « Glove : Global vectors for word representation ». In : *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 2014, p. 1532-1543.

[166] Rute PEREIRA, Jorge OLIVEIRA et Mário SOUSA. « Bioinformatics and Computational Tools for Next-Generation Sequencing Analysis in Clinical Genetics ». In : *Journal of Clinical Medicine* 9.1 (2020). ISSN : 2077-0383. DOI : 10.3390/jcm9010132.

[167] Matthew E PETERS et al. « Deep contextualized word representations ». In : *arXiv preprint arXiv :1802.05365* (2018).

[168] Muller PIERRE-ALAIN et Nathalie GAERTNER. *Modelisation Objet Avec Uml - Muller - 2ème édition - Librairie Eyrolles*. fr. Eyrolles., 2002. (Visité le 07/04/2019).

[169] The PLANT et Ontology CONSORTIUM. « The Plant Ontology Consortium and plant ontologies. » In : *Comparative and functional genomics* 3.2 (2002). Citation Key : Plant2002, p. 137-42. ISSN : 1531-6912. DOI : 10.1002/cfg.154.

[170] Ryan POPLIN et al. « A universal SNP and small-indel variant caller using deep neural networks ». In : *Nature Biotechnology* 36.10 (nov. 2018), p. 983-987. ISSN : 1546-1696. DOI : 10.1038/nbt.4235. URL : https://doi.org/10.1038/nbt.4235.

[171] Changqin QUAN, Meng WANG et Fuji REN. « An Unsupervised Text Mining Method for Relation Extraction from Biomedical Literature ». en. In : *PLOS ONE* 9.7 (juil. 2014). Publisher : Public Library of Science, e102039. ISSN : 1932-6203. DOI : 10.1371/journal.pone.0102039. URL : https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0102039 (visité le 20/04/2020).

[172] L.R. RABINER. « A tutorial on hidden Markov models and selected applications in speech recognition ». In : *Proceedings of the IEEE* 77.2 (1989), p. 257-286.

[173] Aditya RAO et al. « Phenotype-driven gene prioritization for rare diseases using graph convolution on heterogeneous networks ». eng. In : *BMC medical genomics* 11.1 (juil. 2018), p. 57-57. ISSN : 1755-8794. DOI : 10.1186/s12920-018-0372-8.

[174] N REDASCHI et THE UNIPROT CONSORTIUM. « Uniprot in RDF : Tackling data integration and distributed annotation with the semantic web ». In : *Nature Prec* (2009).

[175] Laurens RIETVELD et Rinke HOEKSTRA. « The YASGUI Family of SPARQL Clients ». In : *Semantic Web Journal* (2015). Citation Key : Rietveld2015YASGUI.

[176] Daniel J. RIGDEN et Xosé M. FERNÁNDEZ. « The 26th annual Nucleic Acids Research database issue and Molecular Biology Database Collection ». eng. In : *Nucleic Acids Research* 47.D1 (jan. 2019), p. D1-D7. ISSN : 1362-4962. DOI : 10.1093/nar/gky1267.

[177] Petar RISTOSKI et Heiko PAULHEIM. « RDF2Vec : RDF Graph Embeddings for Data Mining. » In : *The Semantic Web - ISWC 2016 - 15th International Semantic Web Conference, Kobe, Japan, October 17-21, 2016, Proceedings, Part I*. 2016, p. 498-514. DOI : 10.1007/978-3-319-46523-4_30.

[178] Tim ROCKTÄSCHEL, Michael WEIDLICH et Ulf LESER. « ChemSpot : a hybrid system for chemical named entity recognition ». In : *Bioinformatics* 28.12 (2012), p. 1633-1640.

[179] Mariano RODRIGUEZ-MURO. « Query Rewriting and Optimisation with Database Dependencies in Ontop ». In : (). Citation Key : Rodriguez-Muro.

[180] Marcos Mart\'\inez ROMERO et al. « NCBO Ontology Recommender 2.0 : An Enhanced Approach for Biomedical Ontology Recommendation ». In : *CoRR* abs/1611.0 (2016).

[181] Mathieu ROUARD et al. « GreenPhylDB v2.0 : comparative and functional genomics in plants. » In : *Nucleic acids research* 39.Database issue (2011), p. D1095-102. ISSN : 1362-4962. DOI : 10.1093/nar/gkq811.

[182] L.-A. SANDERSON et al. « Tripal v1.1 : a standards-based toolkit for construction of online genetic and genomic databases ». In : *Database* 2013.0 (oct. 2013), bat075-bat075. ISSN : 1758-0463. DOI : 10.1093/database/bat075. (Visité le 21/05/2018).

[183] Susanna-Assunta SANSONE et al. « FAIRsharing as a community approach to standards, repositories and policies ». In : *Nature Biotechnology* 37.4 (avr. 2019), p. 358-367. ISSN : 1546-1696. DOI : 10.1038/s41587-019-0080-8. URL : https://doi.org/10.1038/s41587-019-0080-8.

[184] André SCHUMACHER et al. « SeqPig : Simple and scalable scripting for large sequencing data sets in hadoop ». In : *Bioinformatics* 30.1 (2014). arXiv : 1307.2331 ISBN : 1367-4811 (Electronic)\r1367-4803 (Linking), p. 119-120. ISSN : 13674803. DOI : 10.1093/bioinformatics/btt601.

[185] Dominik SEELOW, Jana Marie SCHWARZ et Markus SCHUELKE. « GeneDistiller–distilling candidate genes from linkage intervals ». eng. In : *PloS one* 3.12 (déc. 2008), e3874-e3874. ISSN : 1932-6203. DOI : 10.1371/journal.pone.0003874.

[186] Isabel SEGURA-BEDMAR, Víctor SUÁREZ-PANIAGUA et Paloma MARTÍNEZ. « Combining Conditional Random Fields and Word Embeddings for the CHEMDNER-patents task ». In : *Proceedings of the fifth BioCreative challenge evaluation workshop* (2015), p. 90-93.

[187] Guilhem SEMPÉRÉ et al. « Gigwa-Genotype investigator for genome-wide analyses. » In : *GigaScience* 5 (2016), p. 25.

[188] Guilhem SEMPÉRÉ et al. « Gigwa v2—Extended and improved genotype investigator ». en. In : *GigaScience* 8.5 (mai 2019). Publisher : Oxford Academic. DOI : 10.1093/gigascience/giz051. URL : https://academic.oup.com/gigascience/article/8/5/giz051/5488103 (visité le 04/05/2020).

[189] Juan F SEQUEDA et al. « Survey of Directly Mapping SQL Databases to the Semantic Web ». In : (2011), p. 1-33.

[190] B. SETTLES. « ABNER : an open source tool for automatically tagging genes, proteins and other entity names in text ». In : *Bioinformatics* 21.14 (2005), p. 3191-3192.

[191] Burr SETTLES. « Biomedical named entity recognition using conditional random fields and rich feature sets ». In : *International Joint Workshop on Natural Language Processing in Biomedicine and its Applications* (2004), p. 104-107.

[192] Sohrab P SHAH et al. « Atlas - a data warehouse for integrative bioinformatics. » In : *BMC Bioinformatics* 6 (2005), p. 34. DOI : 10.1186/1471-2105-6-34.

[193] Brendan SHEEHAN et al. « A relation based measure of semantic similarity for Gene Ontology annotations. » In : *BMC bioinformatics* 9 (jan. 2008), p. 468. ISSN : 1471-2105. DOI : 10.1186/1471-2105-9-468. (Visité le 29/06/2012).

[194] U Martin SINGH-BLOM et al. « Prediction and validation of gene-disease associations using methods inspired by social network analyses ». eng. In : *PloS one* 8.5 (mai 2013), e58977-e58977. ISSN : 1932-6203. DOI : 10.1371/journal.pone.0058977.

[195] Fatima Zohra SMAILI, Robert HOEHNDORF et Xin GAO. « Onto2Vec : joint vector-based representation of biological entities and their ontology-based annotations ». In : *Bioinformatics* 34.13 (juin 2018), p. i52-i60. ISSN : 1367-4803. DOI : 10.1093/bioinformatics/bty259. (Visité le 06/04/2019).

[196] Fatima Zohra SMAILI, Robert HOEHNDORF et Xin GAO. « OPA2Vec : combining formal and informal content of biomedical ontologies to improve similarity-based prediction ». In : (nov. 2018). DOI : 10.1093/bioinformatics/bty933. (Visité le 06/04/2019).

[197] Damian SMEDLEY et al. « BioMart – biological queries made easy ». In : *BMC Genomics* 10.1 (2009). ISBN : 1471-2164 (Electronic)\r1471-2164 (Linking), p. 22. ISSN : 1471-2164. DOI : 10.1186/1471-2164-10-22.

[198] Barry SMITH et al. « The OBO Foundry : coordinated evolution of ontologies to support biomedical data integration ». In : *Nat Biotech* 25.11 (2007), p. 1251-1255.

[199] Richard N SMITH et al. « InterMine : a flexible data warehouse system for the integration and analysis of heterogeneous biological data. » In : *Bioinformatics (Oxford, England)* 28.23 (déc. 2012). Publisher : Oxford University Press, p. 3163-5. ISSN : 1367-4811. DOI : 10.1093/bioinformatics/bts577.

[200] Das SOURIPRIYA, Sundara SEEMA et Cyganiak RICHARD. *R2RML : RDB to RDF Mapping Language*.

[201] R. STEVENS et al. « TAMBIS : transparent access to multiple bioinformatics information sources. » In : *Bioinformatics* 16.2 (fév. 2000), p. 184-185.

[202] Danai SYMEONIDOU et al. « SAKey : Scalable Almost Key Discovery in RDF Data ». In : *The Semantic Web - ISWC 2014 - 13th International Semantic Web Conference, Riva del Garda, Italy, October 19-23, 2014. Proceedings, Part I*. 2014, p. 33-49. DOI : 10.1007/978-3-319-11964-9_3. URL : https://doi.org/10.1007/978-3-319-11964-9_3.

[203] Jan TAUBERT. « ONDEX - a data integration framework for the life sciences ». eng. In : (2011). (Visité le 26/03/2019).

[204] Jan TAUBERT et al. « Ondex Web : Web-based visualization and exploration of heterogeneous biological networks ». In : *Bioinformatics* 30.7 (2014), p. 1034-1035. ISSN : 14602059. DOI : 10.1093/bioinformatics/btt740.

[205] Ronald C TAYLOR. « An overview of the Hadoop/MapReduce/HBase framework and its current applications in bioinformatics ». In : *BMC Bioinformatics* 11.Suppl 12 (2010), S1. ISSN : 1471-2105. DOI : 10.1186/1471-2105-11-S12-S1.

[206] Marcela K. TELLO-RUIZ et al. « Gramene 2018 : Unifying comparative genomics and pathway resources for plant research ». In : *Nucleic Acids Research* (2018). ISSN : 13624962. DOI : 10.1093/nar/gkx1111.

[207] THE GENE ONTOLOGY CONSORTIUM. « Gene Ontology Consortium : going forward. » In : *Nucleic acids research* 43.D1 (2014), p. D1049-1056. ISSN : 1362-4962. DOI : 10.1093/nar/gku1179.

[208] THE GENE ONTOLOGY CONSORTIUM. « Expansion of the Gene Ontology knowledgebase and resources ». In : *Nucleic acids research* 45.D1 (jan. 2017), p. D331-D338. ISSN : 1362-4962. DOI : 10.1093/nar/gkw1108. URL : https://pubmed.ncbi.nlm.nih.gov/27899567.

[209] Syed Hamid TIRMIZI, Juan SEQUEDA et Daniel MIRANKER. « Translating SQL Applications to the Semantic Web ». In : (2008), p. 450-464.

[210] Syed Hamid TIRMIZI et al. « Mapping between the OBO and OWL ontology languages. » In : *J. Biomedical Semantics* 2 (2011).

[211] Anthony TOMASIC, Louiqa RASCHID et Patrick VALDURIEZ. « Scaling Access to Heterogeneous Data Sources with DISCO ». In : *Knowledge and Data Engineering* 10.5 (1998), p. 808-823.

[212] Anne TOULET, Vincent EMONET et Clement JONQUET. « Modele de metadonnees dans un portail d'ontologies ». In : *JFO : Journ{é}es Francophones sur les Ontologies*. Bordeaux, France, 2016.

[213] Léon Charles TRANCHEVENT et al. « Candidate gene prioritization with Endeavour ». In : *Nucleic acids research* 44.W1 (2016), W117-W121.

[214] Léon-Charles TRANCHEVENT et al. « Kernel-based data fusion for gene prioritization ». In : *Bioinformatics* 23.13 (juil. 2007), p. i125-i132. ISSN : 1367-4803. DOI : 10.1093/bioinformatics/btm187. (Visité le 06/04/2019).

[215] Rashmi TRIPATHI et al. « Next-generation sequencing revolution through big data analytics ». In : *Frontiers in Life Science* 9.2 (avr. 2016). Publisher : Taylor & Francis, p. 119-149. ISSN : 2155-3769. DOI : 10.1080/21553769.2016.1178180. URL : https://doi.org/10.1080/21553769.2016.1178180.

[216] Silke TRISSL et al. « Columba : an integrated database of proteins, structures, and annotations. » In : *BMC Bioinformatics* 6 (2005), p. 81. DOI : 10.1186/1471-2105-6-81.

[217] Thoralf TÖPEL et al. « BioDWH : a data warehouse kit for life science data integration ». eng. In : *Journal of Integrative Bioinformatics* 5.2 (août 2008). ISSN : 1613-4516. DOI : 10.2390/biecoll-jib-2008-93.

[218] Özlem UZUNER et al. « 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. » In : *Journal of the American Medical Informatics Association : JAMIA* 18.5 (2011), p. 552-6.

[219] Alfonso VALENCIA. « Search and retrieve. Large-scale data generation is becoming increasingly important in biological research. But how good are the tools to make sense of the data ? » eng. In : *EMBO reports* 3.5 (mai 2002), p. 396-400. ISSN : 1469-221X. DOI : 10.1093/embo-reports/kvf104.

[220] Aravind VENKATESAN et al. « Agronomic Linked Data (AgroLD) : A knowledge-based system to enable integrative biology in agronomy ». In : *PLOS ONE* 13.11 (2018), p. 1-17. DOI : 10.1371/journal.pone.0198270.

[221] Samart WANCHANA et al. « The Generation Challenge Programme comparative plant stress-responsive gene catalogue ». In : *Nucleic Acids Research* 36.Database issue (2008), p. D943-946.

[222] Hei-Chia WANG et al. « Inference of transcriptional regulatory network by bootstrapping patterns ». In : *Bioinformatics* 27.10 (mar. 2011), p. 1422-1428. ISSN : 1367-4803. DOI : 10.1093/bioinformatics/btr155. URL : https://doi.org/10.1093/bioinformatics/btr155.

[223] Wensheng WANG et al. « Genomic variation in 3,010 diverse accessions of Asian cultivated rice ». In : *Nature* 557.7703 (mai 2018). Publisher : Nature Publishing Group, p. 43-49. ISSN : 0028-0836. DOI : 10.1038/s41586-018-0063-9. (Visité le 23/05/2018).

[224] Xuan WANG et al. « Cross-type biomedical named entity recognition with deep multi-task learning ». eng. In : *Bioinformatics (Oxford, England)* 35.10 (mai 2019), p. 1745-1752. ISSN : 1367-4811. DOI : 10.1093/bioinformatics/bty869.

[225] Patricia L WHETZEL et al. « The MGED ontology : A Resource for semantics-based description of microarray experiments ». In : 22.7 (2006), p. 866-873.

[226] Gio WIEDERHOLD et Michael R. GENESERETH. « The basis for mediation ». In : *Proc. COOPIS* (1995). (Visité le 02/02/2013).

[227] Gio WIEDERHOLD et Michael R. GENESERETH. « The Conceptual Basis for Mediation Services ». In : *IEEE Expert* 12.5 (1997), p. 38-47.

[228] Mark WILKINSON et al. « BioMOBY successfully integrates distributed heterogeneous bioinformatics Web Services. The PlaNet exemplar case ». In : *Plant Physiology* 138.1 (2005), p. 5-17.

[229] Mark D WILKINSON et Matthew LINKS. « BioMOBY : an open source biological web services proposal. » In : *Briefings in Bioinformatics* 3.4 (2002), p. 331-341.

[230] Mark D. WILKINSON et al. « The FAIR Guiding Principles for scientific data management and stewardship ». In : *Scientific Data* 3 (2016), p. 160018.

[231] Antony J. WILLIAMS et al. *Open PHACTS : Semantic interoperability for drug discovery*. 2012.

[232] Rod A. WING et al. « The Oryza Map Alignment Project : The Golden Path to Unlocking the Genetic Potential of Wild Rice Species ». In : *Plant Molecular Biology* 59.1 (sept. 2005), p. 53-62. ISSN : 1573-5028. DOI : 10.1007/s11103-004-6237-x.

[233] Julien WOLLBRETT et al. « Clever generation of rich SPARQL queries from annotated relational schema : application to Semantic Web Service creation for biological databases ». In : *BMC bioinformatics* 14.1 (2013), p. 126-141.

[234] Zhihao XIA et al. « DeeReCT-PolyA : a robust and generic deep learning method for PAS identification ». In : (nov. 2018). DOI : 10.1093/bioinformatics/bty991. (Visité le 06/04/2019).

[235] Wonjin YOON et al. « CollaboNet : collaboration of deep neural networks for biomedical named entity recognition ». In : *BMC bioinformatics* 20.10 (2019), p. 249.

[236] Haiyuan YU, Natali GULBAHCE et Xiujuan WANG. « Network-based methods for human disease gene prediction ». In : *Briefings in Functional Genomics* 10.5 (juil. 2011), p. 280-293. ISSN : 2041-2649. DOI : 10.1093/bfgp/elr024. (Visité le 06/04/2019).

[237] Pooya ZAKERI et al. « Gene prioritization using Bayesian matrix factorization with genomic and phenotypic side information ». eng. In : *Bioinformatics (Oxford, England)* 34.13 (juil. 2018), p. i447-i456. ISSN : 1367-4811. DOI : 10.1093/bioinformatics/bty289.

[238] Daojian ZENG et al. « Relation Classification via Convolutional Deep Neural Network ». In : *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics : Technical Papers*. Dublin, Ireland : Dublin City University et Association for Computational Linguistics, août 2014, p. 2335-2344. URL : https://www.aclweb.org/anthology/C14-1220.

[239] Wei ZHENG et al. « An attention-based effective neural model for drug-drug interactions extraction ». eng. In : *BMC bioinformatics* 18.1 (oct. 2017), p. 445. ISSN : 1471-2105. DOI : 10.1186/s12859-017-1855-x.

[240] Deyu ZHOU et Yulan HE. « Extracting interactions between proteins from the literature ». In : *Journal of Biomedical Informatics* 41.2 (avr. 2008), p. 393-407. ISSN : 1532-0464. DOI : 10.1016/j.jbi.2007.11.008. URL : http://www.sciencedirect.com/science/article/pii/S1532046407001451 (visité le 17/04/2020).

[241] Huiwei ZHOU et al. « Exploiting syntactic and semantics information for chemical-disease relation extraction ». eng. In : *Database : The Journal of Biological Databases and Curation* 2016 (2016). ISSN : 1758-0463. DOI : 10.1093/database/baw048.

[242] Huiwei ZHOU et al. « Knowledge-guided convolutional networks for chemical-disease relation extraction ». In : *BMC Bioinformatics* 20 (mai 2019). ISSN : 1471-2105. DOI : 10.1186/s12859-019-2873-7. URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6528333/ (visité le 20/04/2020).

[243] Huiwei ZHOU et al. « Leveraging prior knowledge for protein–protein interaction extraction with memory network ». In : *Database : The Journal of Biological Databases and Curation* 2018 (juil. 2018). ISSN : 1758-0463. DOI : 10.1093/database/bay071. URL : https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6047414/.

[244] Pinglei ZHOU, David EMMERT et Peili ZHANG. « Using Chado to store genome annotation data ». In : *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al* Chapter 9 (jan. 2006), Unit 9.6. ISSN : 1934-340X. DOI : 10.1002/0471250953.bi0906s12. (Visité le 01/04/2010).

[245] Marinka ZITNIK, Monica AGRAWAL et Jure LESKOVEC. « Modeling polypharmacy side effects with graph convolutional networks ». eng. In : *Bioinformatics (Oxford, England)* 34.13 (juil. 2018), p. i457-i466. ISSN : 1367-4811. DOI : 10.1093/bioinformatics/bty294.

[246] Marinka ŽITNIK et al. « Gene Prioritization by Compressive Data Fusion and Chaining ». eng. In : *PLoS computational biology* 11.10 (oct. 2015), e1004552-e1004552. ISSN : 1553-7358. DOI : 10.1371/journal.pcbi.1004552.

Table des figures

1.1 Plan général . . . 3
1.2 Schéma général du parcours . . . 5

2.1 Schéma général du projet de recherche . . . 11

3.1 Le dogme central de la biologie moléculaire . . . 13
3.2 Différentes échelles de la régulation de l'expression des gènes conduisant à un phénotype . . . 15
3.3 Mécanisme d'ouverture du nucléosome . . . 16
3.4 Arbre Phylogénétique du genre Oryza . . . 18
3.5 Analyse GWAS réalisée pour la longueur du grain (GRLT) chez Oryza sativa . . . 20
3.6 Évolution des systèmes d'information en parallèle des méthodes biologiques . . . 21
3.7 Représentation d'un triplet RDF (sujet, prédicat, objet) . . . 25
3.8 Représentation d'un schéma RDFS . . . 26

4.1 Schéma conceptuel montrant une relation d'héritage. . . . 40
4.2 Graphe représentant la vue RDF enrichie du schéma de la base de données utilisée. . . 41
4.3 Utilisation de l'enrichissement sémantique dans le parcours de graphe. . . . 42
4.4 Processus d'annotation sémantique entre AgroPortal et AgroLD. Dans une première étape AgroLD utilise le Recommender d'AgroPortal pour identifier les ontologies nécessaires pour annoter le texte. Puis, les éléments du texte sont envoyés à l'API d'Annotator avec des paramètres correspondants dont les ontologies sélectionnées. L'Annotator renvoie les résultats qui sont intégrés dans le pipeline d'annotation. . . . 50
4.5 Résultat du Quick Search . . . 51
4.6 Résultat de SPARQLEditor . . . 52
4.7 Exemple de requête SPARQL - Q1 . . . 53
4.8 Exemple de requête SPARQL - Q2 avec Property Path . . . 54
4.9 Résultat du Explore Relationships . . . 55
4.10 Résultat de l'Advanced Search . . . 55
4.11 Vue générale de notre classification SQR . . . 56

5.1 Le modèle LSTM-CRF . . . 65
5.2 Le schéma du Distiller . . . 66
5.3 Workflow général du processus de liage de données . . . 67
5.4 Un exemple de problème de liage de données rencontré dans AgroLD . . . 68
5.5 Schéma général d'un exemple de liage de données . . . 69

Liste des tableaux

1.1 Sélection de projets financés. La majorité des budgets obtenus a été allouée à l'IRD. Le premier budget entre parenthèse correspond à la somme perçue alors que le budget global est indiqué à la suite. 6
1.2 Sélection d'encadrements réalisés. 7

4.1 Les espèces et les sources de données intégrées dans AgroLD. 51

5.1 Résultats des performances des approches LSTM 66
5.2 Résultats des performances des approches OGER 67

Quatrième partie

Sélection de publications

Annexe A

Sélection de publications

Les publications suivantes sont incluses dans le manuscript.

A.1 Journal

1. Sempéré G, Pétel A, Rouard M, Frouin J, Hueber Y, De Bellis F, **Larmande P**. *Gigwa v2 – Extended and improved genotype investigator.* GigaScience, Volume 8, Issue 5, May 2019, giz051 Impact Factor : 7.31
2. Venkatesan A., Tagny G., El Hassouni N., Chentli I., Guignon V., Jonquet C., Ruiz M., and **Larmande P.** *Agronomic Linked Data (AgroLD) : a Knowledge-based System to Enable Integrative Biology in Agronomy.* PLoS ONE 13(11) : e0198270. 2018. Impact Factor : 2.766
3. Jonquet C, Toulet A, Arnaud E, Aubin E, Dzalé-Yeumo E, Emonet V, Graybeal J, Laporte M-A, Musen M, Pesce V, **Larmande P**. *AgroPortal : an ontology repository for agronomy.* Comput. Electron. Agric. 2018 ; 144 :126–143 Impact Factor : 2.201
4. Sempéré G, Philippe F, Dereeper A, Ruiz M, Sarah G, **Larmande P**. *Gigwa—Genotype investigator for genome- wide analyses.* Gigascience. 2016. 5 :25. Impact Factor : 7.31
5. Wollbrett J, **Larmande P**, de Lamotte F, Ruiz M. *Clever generation of rich SPARQL queries from annotated relational schema : application to Semantic Web Service creation for biological databases.* BMC Bioinformatics. 2013. 14 :126-141.Impact Factor : 2.435

GigaScience, 8, 2019, 1–9

doi: 10.1093/gigascience/giz051
Technical Note

TECHNICAL NOTE

Gigwa v2—Extended and improved genotype investigator

Guilhem Sempéré [1,2,3,*], Adrien Pétel [2,4], Mathieu Rouard [2,5], Julien Frouin [6,7], Yann Hueber [2,5], Fabien De Bellis [6,7] and Pierre Larmande [2,4]

[1]Centre de coopération Internationale en Recherche Agronomique pour le Développement (CIRAD), UMR INTERTRYP, F-34398 Montpellier, France , [2]South Green Bioinformatics Platform, Bioversity, CIRAD, Institut National de la Recherche Agronomique (INRA), IRD, Montpellier, France, [3]INTERTRYP, Univ Montpellier, CIRAD, Institut de Recherche pour le Développpement (IRD), Montpellier, France, [4]DIADE, Univ Montpellier, IRD, 911 Avenue Agropolis, 34394 Montpellier, France, [5]Bioversity International, Parc Scientifique Agropolis II, 34397 Montpellier Cedex 5, France, [6]CIRAD, UMR AGAP, F-34398 Montpellier, France and [7]AGAP, Univ Montpellier, CIRAD, INRA, Institut national d'études supérieures agronomiques de Montpellier (Montpellier SupAgro), Montpellier, France

*Correspondence address. Guilhem Sempéré, TA A-17/G ,Campus International de Baillarguet , 34398 Montpellier Cedex 5 ,France E-mail: guilhem.sempere@cirad.fr http://orcid.org/0000-0001-7429-2091

Abstract

Background: The study of genetic variations is the basis of many research domains in biology. From genome structure to population dynamics, many applications involve the use of genetic variants. The advent of next-generation sequencing technologies led to such a flood of data that the daily work of scientists is often more focused on data management than data analysis. This mass of genotyping data poses several computational challenges in terms of storage, search, sharing, analysis, and visualization. While existing tools try to solve these challenges, few of them offer a comprehensive and scalable solution. **Results:** Gigwa v2 is an easy-to-use, species-agnostic web application for managing and exploring high-density genotyping data. It can handle multiple databases and may be installed on a local computer or deployed as an online data portal. It supports various standard import and export formats, provides advanced filtering options, and offers means to visualize density charts or push selected data into various stand-alone or online tools. It implements 2 standard RESTful application programming interfaces, GA4GH, which is health-oriented, and BrAPI, which is breeding-oriented, thus offering wide possibilities of interaction with third-party applications. The project home page provides a list of live instances allowing users to test the system on public data (or reasonably sized user-provided data). **Conclusions:** This new version of Gigwa provides a more intuitive and more powerful way to explore large amounts of genotyping data by offering a scalable solution to search for genotype patterns, functional annotations, or more complex filtering. Furthermore, its user-friendliness and interoperability make it widely accessible to the life science community.

Keywords: genomic variations; VCF; HapMap; PLINK; NoSQL; MongoDB; SNP; indel; web; interoperability; REST; BrAPI; GA4GH

Received: 30 November 2018; **Revised:** 19 February 2019; **Accepted:** 8 April 2019

Background

Nowadays, next-generation sequencing technologies have become a standard tool for many applications in basic biology as well as for medicine and agronomic research. With the decreasing cost of genome sequencing, many laboratories are increasingly adopting genotyping technologies as routine components in their workflows, generating large datasets of genotyping and genome sequence information. Additionally, scientists are also interested in re-using data produced by large international consortia that have performed re-sequencing or high-density genotyping on material from representative, publicly available diversity collections. For instance, the 3000 Rice Genome Project [1] and the 1000 Plants Project [2] provide huge amounts of sequence variation data to search and download through, respectively, their SNP-SEEK [3] or 1001genomes.org portals. Such information is not easy to handle because of its size and its complex structure, both unsupported by standard software such as spreadsheet processors. This kind of data is indeed mostly made available as variant call format (VCF) [4] and often needs to be converted into specific software formats for subsequent analyses (e.g., PLINK [5], Darwin [6], Flapjack [7]). In addition, the tools available to filter data or perform more complex operations are mainly available in the command line. Because these results may contain tens of millions of variants for thousands of samples, scalable and user-friendly solutions need to be offered to the biological community.

We thus developed Gigwa [8] with the aim of providing a system that helps relieve biologists from the burden of technical aspects of variation data manipulation. Gigwa is a web application designed to store large volumes of genotypes (up to tens of billions), initially imported from VCF or other file formats, in a NoSQL database (MongoDB [9]), and to provide a straightforward interface for filtering these data. It makes it possible to navigate within search results, to visualize them in different ways, and to re-export subsets of data into various common formats. In the first version published in 2016, we focused our work on the following important aspects: (i) filtering features that include genotype pattern search, e.g., minor allele frequency (MAF) and missing data ratio to name a few; (ii) storage performance by choosing a NoSQL engine and designing data structure in order to scale with growing dataset sizes and support incremental addition of data into projects; (iii) sharing capabilities, i.e., enabling multiple users to efficiently work on the same datasets without the need to replicate them; and (iv) graphical visualization, which allows either a summarized or detailed view of the dataset contents.

Our experience with biologists operating in various research fields and studying different species helped us improve the application with regard to many aspects. In version 2, we overhauled the graphical interface to improve user experience and visualization features. This new release also integrates a data and user management section to facilitate system administrators' work. We took the evolution of next-generation sequencing and analysis software into account by adding new import and export formats. Gigwa's scaling capacities along with its speed performance were also improved, thus making it able to deal with much larger datasets. Finally, we enabled interoperability with other applications, in particular by implementing standard representational state transfer (REST) application programming interfaces (APIs).

Since the release of Gigwa version 1 [8], the application has been adopted by several institutes, in some cases embedded within information systems like the Musa Germplasm Information System [10], in others deployed as a self-sufficient portal providing convenient access to public data [11]. Feedback was thus collected, suggesting ideas for significant improvement. In this article, we describe the list of newly added features, provide details about software improvements, discuss the benchmarking work done to assess performance progress, and finally present a concrete use case showing the usefulness and efficiency of the application.

Findings

Newly added features

Administration interface

A fully featured administration interface has been implemented, allowing for managing databases, projects, users, and permissions. Thus, it is now possible to manage data visibility and sharing, to suppress existing data, and to grant users read or write permissions on datasets, all with a few mouse-clicks without the need to interact with configuration files as before.

New import functionalities

The first version of Gigwa only supported importing data via specification of an absolute path on the webserver's filesystem. While this method is still supported because it is useful to administrators, new ways of feeding genotyping data into the system have been added:

- By uploading files from the client computer (either using drag and drop or by browsing the filesystem);
- By providing http(s) URLs to online files;
- By specifying the base-URL of a BrAPI [12] v1.1 compliant service that supports genotyping data calls. Indeed, this version embeds a client implementation of BrAPI, which allows users to select a genome map and a study in order to feed a Gigwa project with corresponding genotypes pulled from the BrAPI datasource.

Additionally, the application now allows anonymous users to import genotyping data as temporary databases for filtering purposes. Such datasets are only guaranteed to be maintained online for a limited period. An adjustable size limit can be set for files uploaded by any users, including anonymous ones.

As for import formats, the PLINK (PLINK, RRID:SCR_001757) [5] flat-file standard format is now also supported as input for genotyping data.

Finally, version 2 supports enriching permanent databases by importing metadata as tabulated files for the individuals they refer to. Those metadata facilitate individual selection in the interface based on complementary information beyond the individual identifier (e.g., passport data, traits).

Supported annotation formats

The application is able to take into account functional annotations present in VCF files in order to allow end-users to filter on them. The first version was only able to parse annotations originating from SnpEff (SnpEff, RRID:SCR_005191) [13], whereas version 2 also supports annotations added by VEP (Variant Effect Predictor, RRID:SCR_007931) [14].

New export functionalities
The export features have also been extended as follows:

- The ability to refine the individual list at export time has been added. It is therefore possible to selectively export data relating to a chosen subset of individuals, independently from the one used for filtering variants;
- A new export format was added (.fjzip) for compatibility with the Flapjack [7] software;
- In the case of data files being exported to server, Gigwa v1 provided the means to push this output to a running instance of the IGV (Integrative Genomics Viewer, RRID:SCR_011793) [15] stand-alone software. Version 2 additionally supports pushing it to online tools such as Galaxy (Galaxy, RRID:SCR_006281) [16, 17] or SNiPlay [18]. The list of connected tools can be managed by administrators, and a custom tool can be configured by each end-user.

New filtering capabilities
Gigwa v2 introduces the following new filtering functionalities:

- In the case where individuals are numerous, defining group contents can be fastidious: selection can now be conveniently made by filtering individuals based on imported metadata. The selection made in each group can then be saved in the web browser using the localStorage API [19].
- For data imported from the VCF format, the initial version supported applying thresholds on the per-sample read depth (i.e., DP) and genotype quality (i.e., GQ) fields. The system now provides means to filter genotypes using any genotype-level numeric fields. The availability of such fields is automatically detected and corresponding threshold widgets are dynamically built into the interface when applicable.
- Two groups of individuals can now be defined for filtering. Therefore, any combination of genotype-level filters that was previously possible to express can now be applied to a first group, while a second combination of filters can be applied to a second group.
- One of the genotype patterns that could originally be applied to selected individuals was "All same," resulting in selecting variants for which those individuals all had the same genotype. This option has been made more flexible (thus renamed to "All or mostly the same") and may now be used in conjunction with a similarity ratio, i.e., a percentage defining how many of the selected individuals within the current group shall share the major genotype.
- Thanks to the 2 latter features, the system is now able to discriminate variants with regard to a phenotype. This may be achieved by defining groups according to the phenotype (e.g., resistant vs susceptible), choosing for both the "All or mostly the same" genotype pattern, setting a reasonable similarity ratio, and ticking the "Discriminate groups" checkbox that appears in this situation. This will result in selecting variants where most individuals in each group have the same genotype, that genotype being different between both groups. The usefulness of this functionality is illustrated in the "Gigwa in action" section.

Additional visualization functionalities
On top of the density graph, additional series can now be displayed representing any VCF-defined genotype-level numeric field. The underlying data for these series consist of the given field's cumulated values for a customizable selection of individuals. Thanks to this feature, the density of variants may now be observed with regard to numeric metadata fields such as genotype quality or read depth distribution.

APIs and data interoperability
Much effort has been put into making Gigwa data interoperable:

External, online genome browsers can now be configured for viewing each variant in its genomic context. Administrators have the ability to specify the URL of a default genome browser (e.g., GBrowse [GBrowse, RRID:SCR_006829] [20], JBrowse [JBrowse, RRID:SCR_001004] [21]) per database. End-users may override this default configuration by specifying another tool, thus only affecting their own interface. When such a configuration exists for a database, each variant line in the main browsing interface table features a clickable icon leading to opening the genome browser at the position of the variant so that it can be checked against available tracks.

Moreover, 2 REST APIs have been implemented to automatically provide access to any data imported into the system:

- The GA4GH [22] v0.6.0a5 API. The new graphical user interface (GUI) mentioned above is implemented as a client for this API; i.e., most interaction between Gigwa's client-side and server-side code is performed in compliance with the standards defined by the GA4GH API;
- The BrAPI [12] v1.1 API. Flapjack [7] and BeegMac [23] are examples of clients that are compatible with the data Gigwa serves via BrAPI. The Musa Germplasm Information System [10] also interacts with Gigwa through BrAPI by serving Gigwa-hosted data using a proxy approach.

Both APIs have different purposes and, respectively, work with health-related data and crop-breeding data. One clear overlap between them being the support for sharing genotyping data, we thought it relevant to implement for each API the calls that rely on the type of data held in our system.

Application architecture outline
Figure 1 illustrates the architecture of Gigwa version 2 and summarizes its functionalities.

Software improvements

Description
(i) User-friendly interface.

The entire web interface has been reworked and is now based on Bootstrap V3 [24], which makes it more self-consistent, intuitive, attractive, and cross-browser compatible. In addition, the new GUI brings various enhancements such as support for decimal numbers for filters applying to numeric fields, and the aforementioned facilities for selecting individuals (cf. "New filtering capabilities" section).

Additionally, a web page was added to the interface to allow users to watch process progress when importing genotyping data, or exporting them to a physical file on the webserver (direct downloads require the web browser to remain open at all times and therefore cannot benefit from this feature). Each progress-watching page has a unique URL and can thus be re-opened at any time. This feature is particularly convenient when working with large amounts of data because of the time taken by imports and exports.

(ii) Enhanced performance in terms of query speed.

As a reminder, each search operation is performed via multiple MongoDB aggregation queries targeting evenly sized variant chunks, thus improving response times while allowing progress monitoring.

Table 1. Benchmarking test description

Test No.	Aims	Methods
1	Assess evolution of tool speed performance. Involved Gigwa v1, Gigwa v2, VCFtools v0.1.13 (originally benchmarked) [4], and VCFtools v0.1.16 (latest at assessment time)	Run on configuration 1 using dataset 1 (along with sub-sampled versions, so as to obtain 6 different databases), all with the same number of individuals (i.e., 3,000) but with various numbers of markers. Query was a MAF range between 10% and 30% applied to the first 2,000 individuals
2	(i) Assess performance of latest versions of tools (Gigwa v2 and VCFtools v0.1.16) when simultaneously querying on variant-level (indexed in Gigwa) and genotype-level (unindexed in Gigwa) fields. (ii) Estimate the benefit of migrating to high-performance hardware by monitoring differences in response times between tools	Run on configuration 2 using dataset 1 without its derivatives, sub-sampling now being performed on the fly by restricting the search to a varying list of chromosomes. The query was the same MAF range query as above
3	(i) Test Gigwa v2's suitability for working on very large datasets. (ii) Compare trends with those observed in a small dataset (Test 2)	Run on configuration 2 using dataset 2, sub-sampling being performed on the fly by restricting the search to a varying list of chromosomes. The query was the same MAF range query as above

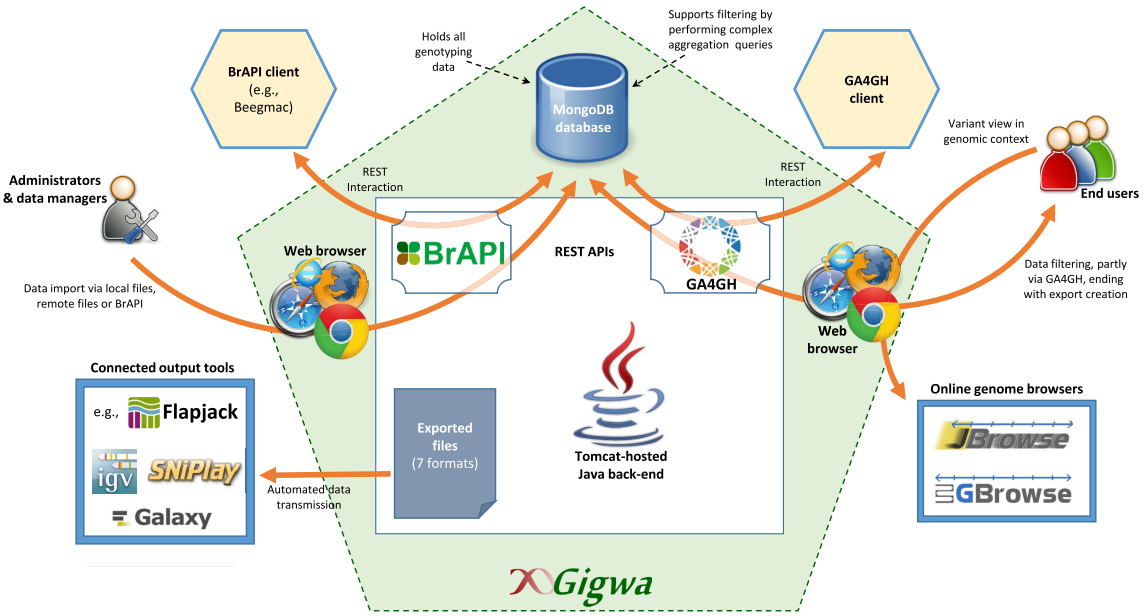


Figure 1: High-level diagram of Gigwa architecture and features.

The data storage structure has also been tuned to optimize speed performance. Gigwa queries consist of combinations of filters that can be split into 2 categories:

- "variant-level" filters (variant type, number of known alleles, sequence and position) applying to indexed fields;
- "genotype-level" filters (all others) applying to non-indexed fields (mainly because a MongoDB collection cannot have >64 indexes, which would be sufficient with only few genotyped individuals).

In version 1, as described by Sempéré et al. [25], indexed fields were held in the "variants" collection only, while unindexed fields (mainly genotype-related) were held in the "variantRunData" collection. Thus, any query involving both types of filters resulted in the following scenario:

- Create a temporary variant collection (subset of the "variants" collection) based on a variant-level query;
- Use the latter collection to restrict genotype-level query target to the variants that matched the variant-level query;
- Update the temporary collection's contents to keep only the variants also matching the genotype-level query.

Although this method worked satisfactorily, it did not scale efficiently with dataset size. Indeed, in the case of a lenient variant-level query, the system would spend much time writing into the temporary collection (copying most of "variants"' contents) and also updating it afterwards (especially when the genotype-level query was stringent).

In version 2, all searchable contents in the variants collection are duplicated into variantRunData, thus allowing all filters to be applied simultaneously by querying a single collection. This data duplication is small and leads to a negligible volume increase that is advantageously compensated by other structure modifications (e.g., removal of empty genotype fields for missing data). This improvement is illustrated in Additional File 1.

In addition, the use of temporary collections has been reduced to a minimum. Previously, any filtering resulted in the creation of temporary data. Thus, only the browsing and exporting of an entire (i.e., unfiltered) database were performed on the main variants collection. With version 2, a temporary collection is created only if a genotype-level query has been submitted. Thus, when the query applies solely to variant-level fields, it is remembered and re-applied to the main variants collection when browsing or exporting data. These indexed queries being extremely fast to execute, responsiveness is not affected even on very large datasets.

Additionally, the JavaScript Object Notation syntax of search queries has been optimized to reduce both the number of operations involved in applying filters and the amount of data processed at each stage of MongoDB's aggregation framework.

Finally, a "multithreading regulation" mechanism was implemented, which adjusts the number of concurrent threads at run time when executing queries. It is based on the database server's live responsiveness and therefore automatically adapts to the current load without taking hardware considerations into account. More detail can be found in Additional File 2.

(iii) Enhanced filtering workflow, improving responsiveness.

When the search button is clicked, depending on the status of the "Enable browse and export" checkbox, the system either builds—and keeps track of—the list of matching variants (find procedure) or simply returns a count value telling how many matching variants were found (count procedure). Each query count is cached as an array of sub-values (1 for each genome chunk), the sum of which equals the query's total result count. These cached values are used when the same query is invoked anytime later; they allow instant response for the count method, and faster response for the find method (thanks to MongoDB's $limit operator, which prevents the aggregation pipeline engine from searching further than the last matching variant in each chunk). In version 1, the count method was always executed prior to the find method, thus almost doubling unnecessarily the execution time when the box was checked. In this situation, version 2 overcomes this problem via a find method that supports a "count at the same time" option. This way, the query is only executed once with a negligible overhead, resulting in much faster display of the results and access to export functionalities.

(iv) Enhanced export and visualization features.

When creating export files, instead of synchronously reading data chunks from the database and writing them to the output stream, we implemented 2 separate processes, one dedicated to reading, the other dedicated to writing, both designed to run concurrently. The reading process was optimized using a multithreading regulation routine as described above.

As for the density visualization functionality, it was improved by making chart zooming dynamic: a new query is now sent to the server each time the zoom level changes, thus always ensuring optimal data resolution.

Benchmarking

We performed benchmarking tests to (i) assess how tools tested in our previous article evolved in terms of speed, (ii) demonstrate the benefit of targeting a genome region when applying a genotype-level filter in Gigwa, and (iii) evaluate our system's capacity to work with very large datasets.

Two hardware configurations were used in this benchmark:

Configuration 1: comparable to the one tested in the original benchmark [8], and essentially used for assessing the progress made since then. It is a Hewlett Packard EliteBook 850 G3 laptop computer with an Intel Core i7–6500U central processing unit (CPU) at 2.50 GHz, 16 GB of random access memory (RAM), and a Samsung PM871 512 GB (6Gbit/s) TLC SSD 850.

Configuration 2: high-performance machine typically suitable to serve as a production environment for MongoDB and thus for Gigwa. We used it to evaluate the performance of the latest software versions running on production hardware, including on large datasets. It is a Dell PowerEdge R640 server based on an Intel Xeon Gold 5122 CPU at 3.60 GHz, 384 GB of RAM, and a 1.92To SAS (12Gbit/s) Toshiba PX05SV SSD.

Two datasets were used in this benchmark:

Dataset 1: dataset tested in the original benchmark, the Old Subset SNP Dataset v0.2.1 (formerly named CoreSNP v2.1) from the 3000 Rice Genomes Project [26], containing genotypes for 3,000 individuals on 365,710 single-nucleotide polymorphisms (SNPs). Its reasonable size (4.4 GB in VCF format) was suitable for experimenting with Configuration 1.

Dataset 2: filtered SNP v1.0 Dataset from the 3000 Rice Genomes Project, containing genotypes for 3,024 individuals on 4,817,964 SNPs (VCF file of 60.4 GB, preliminarily annotated with SnpEff v4.3T).

Because it was demonstrated in the original article that relational database management system–based implementations were suitable only for querying on indexed fields (which Gigwa v1 could do nearly as efficiently) but not on genotype-level information, such solutions were left out in the present work. Therefore, we mostly concentrated on executing queries at the genotype level, especially using the MAF range query, which is among the most CPU-intensive. All Gigwa instances were set up with

MongoDB's WiredTiger storage engine, using the zlib compression level, which had seemed to be the best option in the original tests.

Three different comparison tests were run in this benchmark, which are described in Table 1 and whose results are reported in Fig. 2

Average response times were calculated based on the results provided in Additional File 3.

Looking at Test 1 trends, and considering the results of the original benchmark, the speed difference between VCFtools (VCFtools, RRID:SCR_001235) and Gigwa increased substantially. Because the binaries used in both tests were the same for the versions initially assessed, this difference is due to hardware considerations (the amount of RAM was reduced from 32 to 16 GB) and stems from the fact that Gigwa, being a 3-tier web application, cannot be as lightweight as VCFtools and thus requires more memory to achieve similar performance (cf. Test 2).

The main goal in this test was to compare results tool by tool, thus assessing speed evolution between former and current versions. A substantial speed gain ranging between 18.5% and 36.5% was observed in moving from Gigwa v1 to Gigwa v2. However, rather oddly, a consistent speed loss ranging between 40% and 48.5% was observed in moving from VCFtools v0.1.13 to v0.1.16.

From Test 2 results, a first observation is that using production hardware in which much RAM is available for MongoDB and Tomcat, the difference in speed between tools is far smaller. If we take the full dataset (365,710 variants) as a comparison reference, the Gigwa v2 query takes 3.8 times longer to execute than with VCFtools v0.1.16 for Test 1, whereas for Test 2 it is only 1.36 times slower.

Besides, when targeting a region of the genome, Gigwa takes advantage of its indexing strategy and even responds faster than VCFtools as the given region becomes narrow enough.

Test 3 results demonstrate that Gigwa v2 is able to efficiently handle and search very large datasets (here, >14 billion genotypes) when running on suitable hardware. Also, the trend observed in Test 2 is confirmed here, i.e., targeting a precise genome region for applying a genotype-oriented filter is of great benefit in terms of speed.

Benchmark discussion. The benchmarking work previously performed by Sempéré et al. [8] had shown that, and provided reasons why VCFtools excels in executing genotype-level queries on an entire large dataset. Although equaling its performance in a 3-tier application like the one presented here does not seem feasible for such queries, we thought it relevant to assess the progress made since version 1, still in comparison with VCFtools. This work led to several conclusions: (i) Gigwa v2 performs substantially better than v1 in applying genotype-level queries; (ii) setting up Gigwa on production hardware (with large amounts of RAM) greatly improves its performance; (iii) combining variant-level and genotype-level filters whenever possible is a good way to make the most of Gigwa's indexed fields and can lead it to outperform VCFtools.

In a separate work lying outside the scope of this article, we tested Gigwa v2 configured as a sharded cluster on a single server (Configuration 2). We observed a speed gain within the 20–30% range, which we consider interesting, but we acknowledge the complexity that it induces in terms of application deployment and maintenance. Further investigation would therefore be required to propose best practices in deploying an optimized configuration.

Gigwa in action

In order to demonstrate the user-friendliness of the application, we selected a research study that reported the identification of a major quantitative trait locus (QTL) for sex determination in *Pundamilia* (a genus of cichlid fish), which was achieved by construction of a linkage map [27]. Because the genotype and phenotype files had been made available by the authors [28], it was straightforward to load them into Gigwa, assign all males (144) to group 1 and all females (78) to group 2, and apply a discrimination filter between them, with missing data maximum set to 10% and similarity ratio set to 90% for both groups. By ticking the discrimination filter, we made sure to restrict the results to variants showing a difference between groups.

As shown in Fig. 3, 14 matching variants were found outright on the sole chromosome 10, all but 4 of them concentrating in the 27.53–29.52 megabase region. Findings from the original study indeed indicate that Pun-LG10 "acts as an (evolving) sex chromosome," and that "the QTL region (Bayesian confidence interval) for sex determination in *Pundamilia* is located between 27.8 and 29.7 Mb" (Fig. 3C). Interestingly, by fine-tuning the similarity ratio as a cursor, we noticed that increasing it to 92% narrowed down results to variants exclusively concentrated in the mentioned QTL, while decreasing it to 89% revealed a few variants on unanchored scaffolds that could potentially be interpreted as belonging to Pun-LG10. Besides, we spotted the 2 individuals, 21,321 and 21,327 (Fig. 3B), that were labeled as females but had a male genotype as mentioned by Feulner et al. [27]. This shows that Gigwa can support rapid data exploration in order to provide a valuable indication for similar research studies. Through this example, we demonstrate that our software, although clearly not a replacement for methods such as genome-wide association studies or QTL mapping, provides a means to quickly obtain rough trends regarding the relationships between phenotypes and loci or genotypes, with only a few clicks.

Conclusions

Gigwa v2 is a user-friendly, species-agnostic web application for managing and exploring high-density genotyping data. The software can be installed on a local computer or deployed as a data portal. It supports various standard import and export formats and provides advanced filtering options as well as means to visualize density charts or push selected data into various stand-alone or online tools. It implements 2 standard REST APIs: GA4GH, which is health-oriented, and BrAPI, which is breeding-oriented, thus offering wide possibilities of interaction with other systems.

Once installed, which is done by simply decompressing a zip archive for "stand-alone" users, its management interface obviates the need for any particular computer skills for users to administer, publish, or share their data.

Since its original version, Gigwa's data structure and query syntax have been optimized to a point where its speed performance is comparable to that of state-of-the-art command-line tools when running on production hardware. For instance, this version is able to deal with datasets as large as the 3000 Rice Genomes CoreSNP (genotypes for 3,024 individuals on 4,817,964 SNPs). Current live instances listed at [29] provide access to a range of diverse public datasets [26, 30–33] as well as video demonstrations to facilitate use and adoption.

Gigwa v2 allows for anonymous users to import their own data into temporary databases, thus allowing anyone to test the system on the mentioned live instances, for a limited duration.

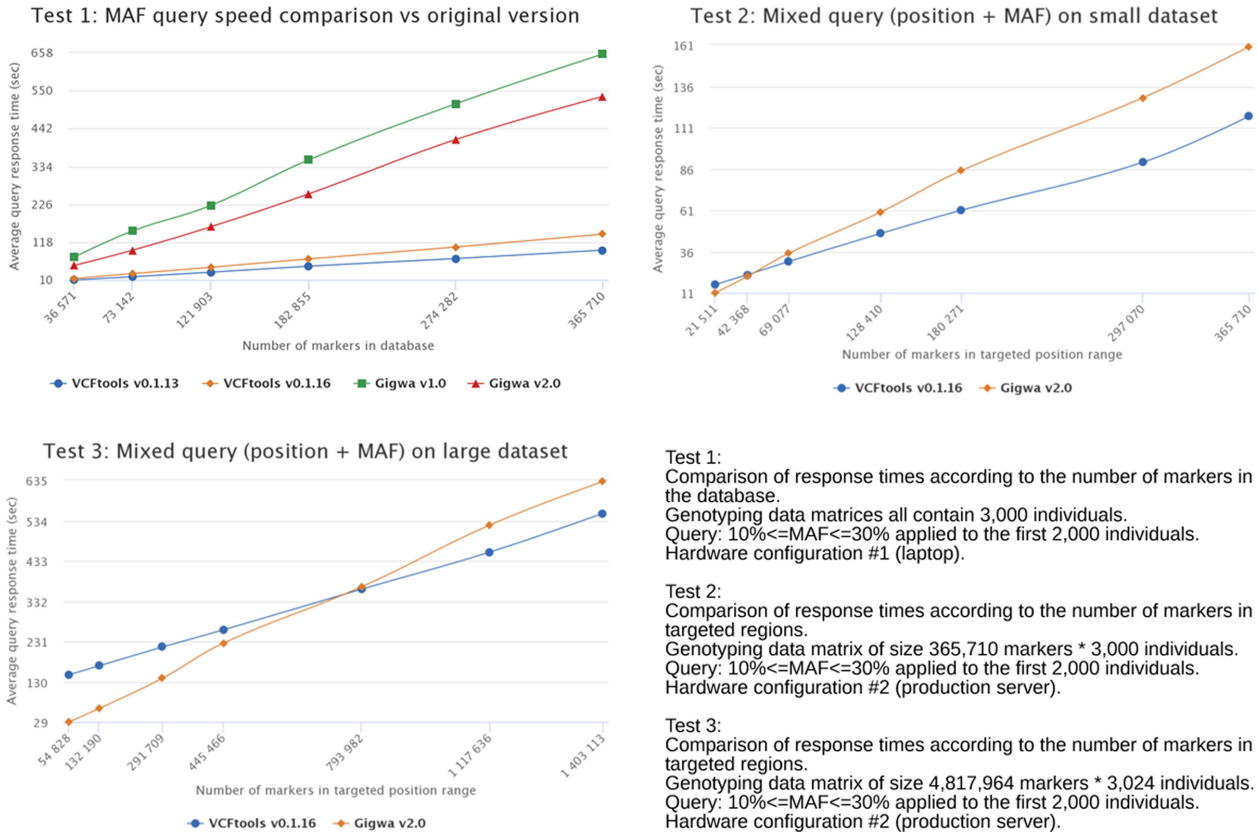


Figure 2: Benchmark results.

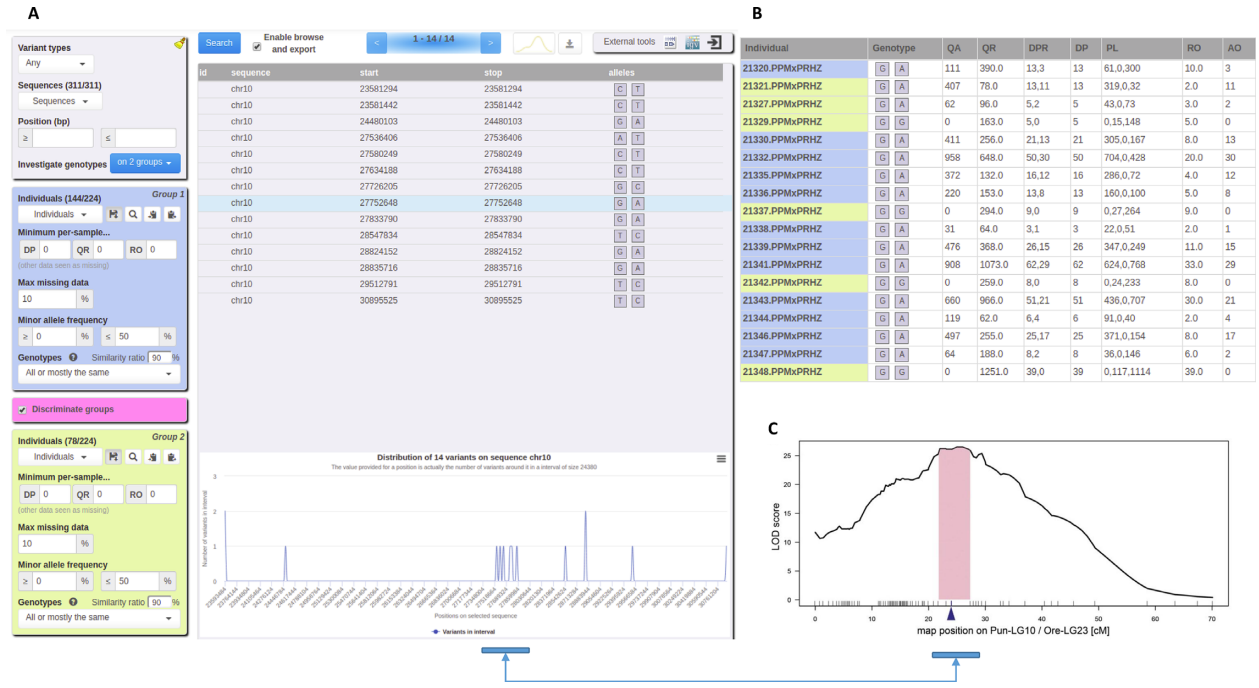


Figure 3: Discriminating variants. A, Filtering parameters and variant distribution. B, A discriminated variant's genotype with complementary information (males in blue, females in yellow). C, Chromosome region as reported by Feulner et al. [27] showing the strongest association with sex determination.

Its filtering functionalities are advanced enough to rapidly obtain an overview of variants discriminating 2 groups of individuals.

The type of data managed by this application being central to many kinds of studies in the genomics field, a wide range of extensions can be envisioned in terms of metadata support, downstream analyses, or visualization. In addition, speed improvements can still be envisioned by means of deep investigation of sharded cluster deployment possibilities.

Availability of supporting source code and requirements

- Project name: Gigwa v2
- Project home page: http://www.southgreen.fr/content/gigwa
- Research Resource Identifier: Gigwa, RRID:SCR_017080
- Operating system(s): Platform-independent
- Programming languages: Java, MongoDB, HTML, Javascript
- Requirements: Java 8 or higher, Tomcat 8 or higher, MongoDB 3.4 or higher
- License: GNU Affero General Public License v3.0
- Restrictions to use for non-academics: None

Availability of supporting data and materials

Gigwa's source code is available in the South Green GitHub repository [34, 35]. Supplementary data, benchmarking material, and installation archives can be found in the *GigaScience* GigaDB repository [36].

Additional files

Additional File 1: Improvement on execution of mixed queries.
Additional File 2: Multithreading regulation explained.
Additional File 3: Detailed benchmark figures.

Abbreviations

API: application programming interface; CPU: central processing unit; GUI: graphical user interface; MAF: minor allele frequency; QTL: quantitative trait locus; RAM: random access memory; REST: representational state transfer; SNP: single-nucleotide polymorphism; VCF: variant call format.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

A.P. implemented the GA4GH service and integrated it into the client-server communication code. A.P. and G.S. designed the new GUI. G.S. implemented all other improvements and additions, optimized the data structure and application speed, and designed and ran the benchmarks. J.F., Y.H., M.R., and F.d.B. helped debugging by deeply testing the system, and suggested new features. G.S., P.L., and M.R. wrote the manuscript. All authors read and approved the final manuscript.

Funding

This work was performed in the frame of the I-SITE MUSE "AdaptGrass project", publicly funded through ANR (the French National Research Agency) under the "Investissements d'avenir" programme with the reference ANR-16-IDEX-0006.

Acknowledgments

The authors thank the South Green Platform team for technical support. We are also grateful to Manuel Ruiz, Stéphanie Sidibe-Bocs, Benjamin Penaud, and Gaëtan Droc for promoting the software and providing new feature ideas, Iain Milne and Gordon Stephen for sharing their Java BrAPI client code, and Jean-Marc Mienville for careful reading that helped improve the manuscript. Finally, we render thanks to Bioversity International, UMR Diade, UMR AGAP, and UMR BGPI for investing in high-performance servers used for hosting public Gigwa instances.

References

1. Wang W, Mauleon R, Hu Z, et al. Genomic variation in 3,010 diverse accessions of Asian cultivated rice. Nature 2018;**557**:43–9.
2. Alonso-Blanco C, Andrade J, Becker C, et al. 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. Cell 2016;**166**:481–91.
3. Alexandrov N, Tai S, Wang W, et al. SNP-Seek database of SNPs derived from 3000 rice genomes. Nucleic Acids Res 2015;**63**:2–6.
4. Danecek P, Auton A, Abecasis G, et al. The variant call format and VCFtools. Bioinformatics 2011;**27**:2156–8.
5. Slifer SH. PLINK: key functions for data analysis. Curr Protoc Hum Genet 2018;**97**:e59.
6. DARwin - Dissimilarity Analysis and Representation for Windows. http://darwin.cirad.fr/. Accessed 21 November 2018.
7. Milne I, Shaw P, Stephen G, et al. Flapjack–graphical genotype visualization. Bioinformatics 2010;**26**:3133–4.
8. Sempéré G, Philippe F, Dereeper A, et al. Gigwa-Genotype investigator for genome-wide analyses. GigaScience 2016;**5**:25.
9. MongoDB. 2015. https://www.mongodb.org/. Accessed 19 December 2015.
10. Ruas M, Guignon V, Sempere G, et al. MGIS: Managing banana (*Musa* spp.) genetic resources information and high-throughput genotyping data. Database (Oxford) 2017;**2017**:, doi: 10.1093/database/bax046.
11. Cubry P, Tranchant-Dubreuil C, Thuillet AC, et al. The rise and fall of African rice cultivation revealed by analysis of 246 new genomes. Curr Biol 2018;**28**:2274–82.e6.
12. Selby Peter, Abbeloos R, Backlund JE, et al. BrAPI - an application programming interface for plant breeding applications. Bioinformatics 2019, doi:10.1093/bioinformatics/btz190.
13. Cingolani P, Platts A, Wang LL, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of *Drosophila melanogaster* strain w 1118; iso-2; iso-3. Fly (Austin) 2012;**6**(2):80–92.
14. McLaren W, Gil L, Hunt SE, et al. The ensembl variant effect predictor. Genome Biol 2016;**17**:122.
15. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative Genomics Viewer (IGV): High-performance genomics data visualization and exploration. Brief Bioinform 2013;**14**:178–92.
16. Goecks J, Nekrutenko A, Taylor J. Galaxy: A comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. Genome Biol 2010;**11**:1–13.
17. Afgan E, Baker D, van den Beek M, et al. The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2016 update. Nucleic Acids Res 2016;**44**(W1):W3–W10.

18. Dereeper A, Homa F, Andres G, et al. SNiPlay3: A web-based application for exploration and large scale analyses of genomic variations. Nucleic Acids Res 2015;**43**:W295–300.
19. Hickson I, Web Storage (Second Edition). https://www.w3.org/TR/webstorage/. Accessed 21 November 2018.
20. Stein LD, Mungall C, Shu S, et al. The generic genome browser: A building block for a model organism system database. Genome Res 2002;**12**:1599–610.
21. Skinner ME, Uzilov AV, Stein LD, et al. JBrowse: A next-generation genome browser. Genome Res 2009;**19**:1630–8.
22. The Global Alliance for Genomics and Health Consortium. GA4GH API. 2017. https://github.com/ga4gh/ga4gh-schemas.Accessed 1 October 2018.
23. Carceller P. beegmac. Github SouthGreen. 2018. https://github.com/SouthGreenPlatform/beegmac. Accessed 1 October 2018.
24. Introduction Bootstrap. http://getbootstrap.com/docs/4.1/getting-started/introduction/. Accessed 1 October 2018.
25. Sempéré G, Moazami-Goudarzi K, Eggen A, et al. WIDDE: A Web-Interfaced next generation database for genetic diversity exploration, with a first application in cattle. BMC Genomics 2015;**16**:940.
26. The 3000 rice genomes project. The 3,000 rice genomes project. GigaScience 2014;**3**:7.
27. Feulner PGD, Schwarzer J, Haesler MP, et al. A dense linkage map of Lake Victoria cichlids improved the *Pundamilia* genome assembly and revealed a major QTL for sex-determination. G3 (Bethesda) 2018;**8**:2411–20.
28. Feulner P, Schwarzer J, Haesler M, et al. Data from: A dense linkage map of Lake Victoria cichlids improved the *Pundamilia* genome assembly and revealed a major QTL for sex-determination. Dryad Digital Repository 2018,doi:10.5061/dryad.59q56g6.
29. Gigwa. http://www.southgreen.fr/content/gigwa. Accessed 1 October 2018.
30. Gibbs RA, Belmont JW, Hardenbol P, et al. The International HapMap Project. Nature 2003;**426**:789–96.
31. Sardos J, Rouard M, Hueber Y, et al. A genome-wide association study on the seedless phenotype in banana (*Musa* spp.) reveals the potential of a selected panel to detect candidate genes in a vegetatively propagated crop. PloS One **11**:e0154448.
32. Nelson J, Wang S, Wu Y, et al. Single-nucleotide polymorphism discovery by high-throughput sequencing in sorghum. BMC Genomics 2011;**12**:352.
33. Soto JC, Ortiz JF, Perlaza-Jiménez L, et al. A genetic map of cassava (*Manihot esculenta* Crantz) with integrated physical mapping of immunity-related genes. BMC Genomics 2015;**16**:190.
34. South Green collaborators. The South Green portal: a comprehensive resource for tropical and Mediterranean crop genomics. Curr Plant Biol 2016;**78**:6–9.
35. South Green Bioinformatic Platform. Gigwa code repository. 2015. https://github.com/SouthGreenPlatform/Gigwa2. Accessed 19 November 2018.
36. Sempéré G, Pétel A, Rouard M, et al. Supporting data for "Gigwa v2—Extended and improved genotype investigator." GigaScience Database 2019. http://dx.doi.org/10.5524/100585.

PLOS ONE

RESEARCH ARTICLE

Agronomic Linked Data (AgroLD): A knowledge-based system to enable integrative biology in agronomy

Aravind Venkatesan[1,2], Gildas Tagny Ngompe[1,2], Nordine El Hassouni[1,3,4], Imene Chentli[1,2], Valentin Guignon[4,5], Clement Jonquet[1,2], Manuel Ruiz[1,3,4,6], Pierre Larmande[1,2,4,7]*

1 Institut de Biologie Computationnelle (IBC), Univ. of Montpellier, Montpellier, France, **2** LIRMM, Univ. of Montpellier & CNRS, Montpellier, France, **3** UMR AGAP, CIRAD, Montpellier, France, **4** South Green Bioinformatics Platform, Montpellier, France, **5** Bioversity International, Montpellier, France, **6** AGAP, Univ. of Montpellier, CIRAD, INRA, INRIA, SupAgro, Montpellier, France, **7** DIADE, IRD, Univ. of Montpellier, Montpellier, France

* pierre.larmande@ird.fr

Citation: Venkatesan A, Tagny Ngompe G, Hassouni NE, Chentli I, Guignon V, Jonquet C, et al. (2018) Agronomic Linked Data (AgroLD): A knowledge-based system to enable integrative biology in agronomy. PLoS ONE 13(11): e0198270. https://doi.org/10.1371/journal.pone.0198270

Editor: Le Zhang, Sichuan University, CHINA

Received: May 14, 2018

Accepted: September 3, 2018

Published: November 30, 2018

Data Availability Statement: The data underlying this study have been uploaded to Zenodo and are accessible using the following DOI: https://doi.org/10.5281/zenodo.1410742.

Funding: This research was supported by the Computational Biology Institute of Montpellier (ANR-11-BINF-0002 - http://www.agence-nationale-recherche.fr/ProjetIA-11-BINF-0002 - project: http://www.ibc-montpellier.fr), the Institut Francais de Bioinformatique (ANR-11-INBS-0013 - http://www.agence-nationale-recherche.fr/ProjetIA-

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 their 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 (Resource Description Format) 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 evaluation results show users appreciate the multiple query modes which support different use cases. AgroLD's 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.

11-INBS-0013 - project: http://www.france-bioinformatique.fr), the Labex Agro (ANR-10-LABX-001-01 - http://www.agence-nationale-recherche.fr/ProjetIA-10-LABX-0001 - project: http://www.agropolis-fondation.fr/) all bypass of the French ANR Investissements d'Avenir program (http://www.agence-nationale-recherche.fr/investissements-d-avenir). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

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 these 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 developers [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.

Materials and methods

Information sources

AgroLD was conceived to accommodate molecular and phenotypic information available on various plant species (see Fig 1). The conceptual framework for the knowledge in AgroLD is based on well-established ontologies: GO, SO, PO, Plant Trait Ontology (TO) and Plant Environment Ontology (EO). Among these PO, TO and EO are currently developed by the Planteome project [29] (http://planteome.org). Furthermore, considering the scope of the effort, we decided to build AgroLD in phases. The current phase (phase I) covers information on genes, proteins, ontology associations, homology predictions, metabolic pathways, plant traits, and germplasm, relevant to the selected species. At this stage, we have incorporated the corresponding information from various databases, such as Gramene [30], UniprotKB [31], Gene Ontology Annotation [32], TropGeneDB [33], OryGenesDB [34], Oryza Tag Line [35], GreenPhylDB [36] and SNiPlay [37]. The selection of these data sources was considered based on

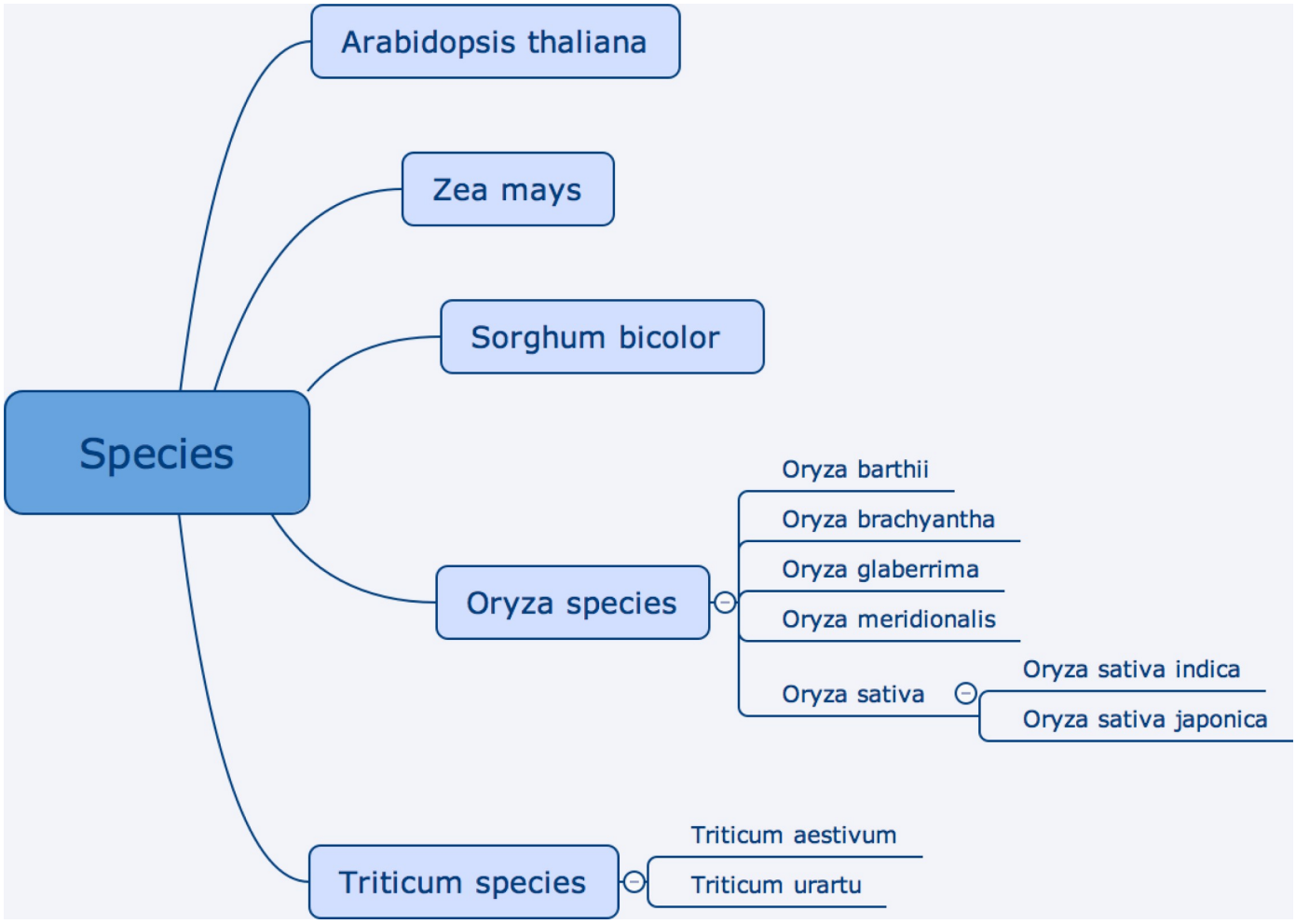


Fig 1. Current plant species included in AgroLD.

https://doi.org/10.1371/journal.pone.0198270.g001

popularity among domain experts such as GOA, Gramene, and complementary information hosted by the local research community, for instance, Oryza Tag Line and GreenPhylDB. Information on the integrated databases can be found in the documentation page (http://www.agrold.org/documentation.jsp). Table 1 provides a break-down of the data sources and the species covered.

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) [38], Generic File Format (GFF3) [39], HapMap [40] and Variant Call Format (VCF) [41]. 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 (https://doi.org/10.5281/zenodo.1294660) is available on GitHub (https://github.com/SouthGreenPlatform/AgroLD).

Table 1. Plant species and data sources in AgroLD.

Data sources	URLs	File format	#tuples	Crops	Ontologies used	#triples produced
GO associations	geneontology.org	GAF	1, 160K	R, W, A, M, S	GO, PO, TO, EO	6, 200K
Gramene	gramene.org	Custom flat file	1, 718K	R, W, M, A, S	GO, PO, TO, EO	4, 600K
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	14, 800K
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	160K
TOTAL						92, 640K

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).

https://doi.org/10.1371/journal.pone.0198270.t001

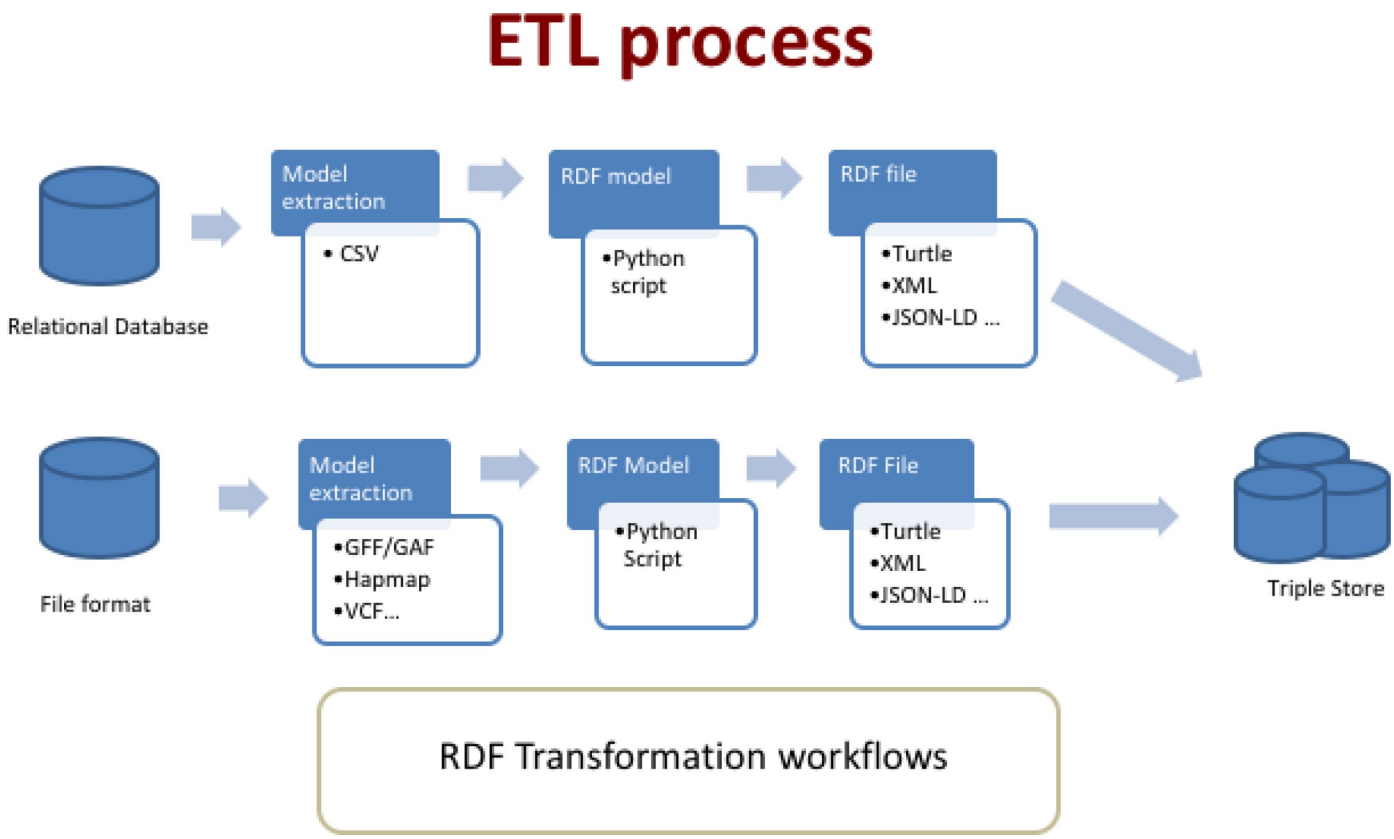


Fig 2. ETL workflow for the various datasets and data formats. The workflow shows two types of process: 1) from relational databases through a CVS file export: in that case, the transformation is tailored for the database model with some Python scripts converters. 2) from standards file formats: in that 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.

https://doi.org/10.1371/journal.pone.0198270.g002

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 used 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 produced additional properties with the corresponding ontologies thus adding 22% additional triples validated manually (see details in Table 1). The OWL versions of the candidate ontologies were directly loaded into the knowledge base but their triples are not counted in the total. We provided in the supplementary file S1 Table, a more comprehensive statistics analysis such as number of triples, classes, entities and properties for each graph stored in the knowledge base.

The RDF graphs are named after the corresponding data sources (protein/qtl ontology annotations being the exception), sharing a common namespace: "http://www.southgreen.fr/agrold/". The entities in the RDF graphs are linked by shared common URIs. As a design principle, we have used URI schemes made available by the sources (e.g., UniprotKB) or by Identifiers.org registry (*http://identifiers.org* - [42]). For instances, proteins from UnitProtKB are identified by the base URI: *http://purl.uniprot.org/uniprot/*; genes incorporated from Gramene/Ensembl plants are identified by the base URI: *http://identifiers.org/ensembl.plant/*. New URIs were minted when not provided by the sources or the by Identifiers.org such as TropGene and OryGenesDB; in such cases the URIs take the form *http://www.southgreen.fr/agrold/**[resource_namespace]/[identifier]***. Furthermore, properties linking the entities took the form: *http://www.southgreen.fr/agrold/vocabulary/**[property]***. An outline of how the RDF graphs are linked is shown in Fig 3. About entity linking, we used the "key-based approach" which is the most common one. It combines the unique identifier/accession number of the entity shared with the community, with the URI basis pattern of the resource. Moreover, we also respected

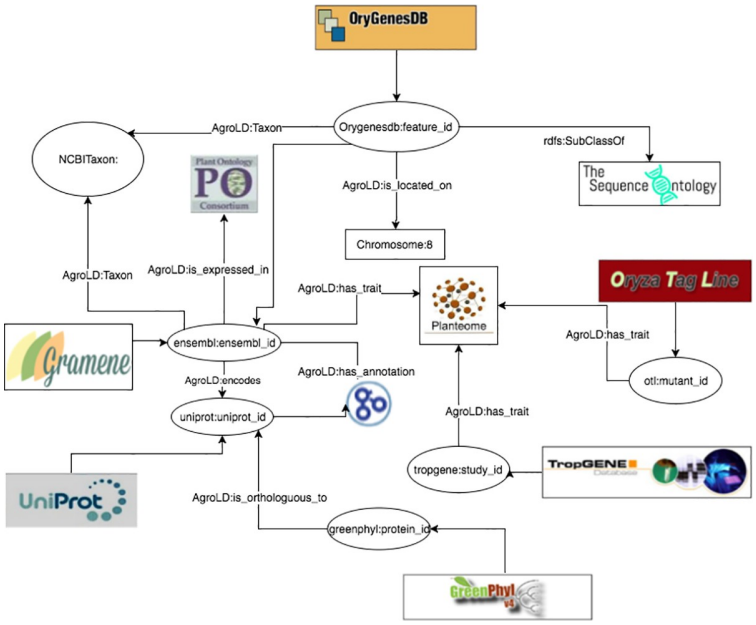


Fig 3. Linking information in AgroLD. The figure illustrates the linking of varies information in AgroLD.

https://doi.org/10.1371/journal.pone.0198270.g003

the "common URI approach" which recommends to use the same URI pattern when the same accession number is used in different datasets. Therefore, defining the same URI for identical entities (represented by identifiers) in different datasets makes it possible to aggregate additional information for this entity. Additionally, we used cross-reference links (represented by identifiers from external datasets) by transforming them into URIs and linked the resource with the predicate "has_dbxref'. This greatly increases the number of outbound links, making AgroLD more integrated with other Linked Open Data. In the future, we will implement a "similarity-based approach" to identify correspondences between entities which have different URIs.

To map the various data types and properties, we developed a lightweight schema (cf. https://github.com/SouthGreenPlatform/AgroLD) that glues classes and properties identified in AgroLD and the corresponding external ontologies. For instance, the class Protein (*http://www.southgreen.fr/agrold/resource/Protein*) is mapped as *owl:equivalentClass* to class polypeptide (*http://purl.obolibrary.org/obo/SO_0000104*) from SO. Similar mappings have been made for properties, e.g., proteins/genes are linked to GO molecular function by the property *http://www.southgreen.fr/agrold/vocabulary/has_function*, which is mapped as *owl:equivalentProperty* to the corresponding Basic Formal Ontology (BFO) term (*http://purl.obolibrary.org/obo/BFO_0000085*). When an equivalent property did not exist, we mapped then to the closest upper level property using *rdfs:subPropertyOf* e.g., the property *has_trait* (*http://www.southgreen.fr/agrold/vocabulary/has_trait*), links proteins to TO terms. It is mapped to a more generic property, *causally related to* in the Relations Ontology [43]. For now, 55 mappings were identified. Furthermore, mappings are both stored side by side with ontologies in AgroPortal, which allows direct links between classes and instances of these classes in AgroLD. For example, the following link will show the external mappings for SO:0000104 (polypeptide) stored in AgroPortal: http://agroportal.lirmm.fr/ontologies/SO/?p=classes&conceptid=http%3A%2F%2Fpurl.obolibrary.org%2Fobo%2FSO_0000104&jump_to_nav=true#mappings. Additionally, classes, properties and resources (e.g., http://www.southgreen.fr/agrold/page/biocyc.pathway/CALVIN-PWY) are dereferenced on a dedicated Pubby server [44]. For details on the graphs, URIs and properties, the reader may refer to AgroLD's documentation (http://www.agrold.org/documentation.jsp).

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;
- *Explore Relationships* visualizer (http://www.agrold.org/relfinder.jsp), which is an implementation of RelFinder [45] 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. Furthermore, other examples are shown in the User Guide available in the supporting information S1 File.

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 [46] and customized it for our system. The SPARQL language is a powerful tool to mine and extract meaningful information from the knowledge base. In the first example of the supplementary S3 File, we compare two queries to answer the question: "Identify wheat proteins that are involved in root development.". While the first one (S3_Q1) using a simple search—which is a direct translation of SQL—with the corresponding id ("GO_0048364", "GO_2000280") shows 73 entries, the second one (S3_Q2) using a property path query (i.e., query the descending class hierarchy for a given trait ontology term) shows 137 entries, thus more than 80% of additional results. In that case, the use of property path algorithm shows the efficiency in retrieving a comprehensive answer. But the SPARQL language performs also very well with complex queries such as: "Retrieve individuals which have positive SNP variant effect identified for proteins associated with a QTL" available in S3_Q3. This type of query involves several datasets and uses graph traversal property of SPARQL to perform the query.

Because SPARQL is hard to handle for non-technical users, the *SPARQL Query Editor* 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.

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.

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

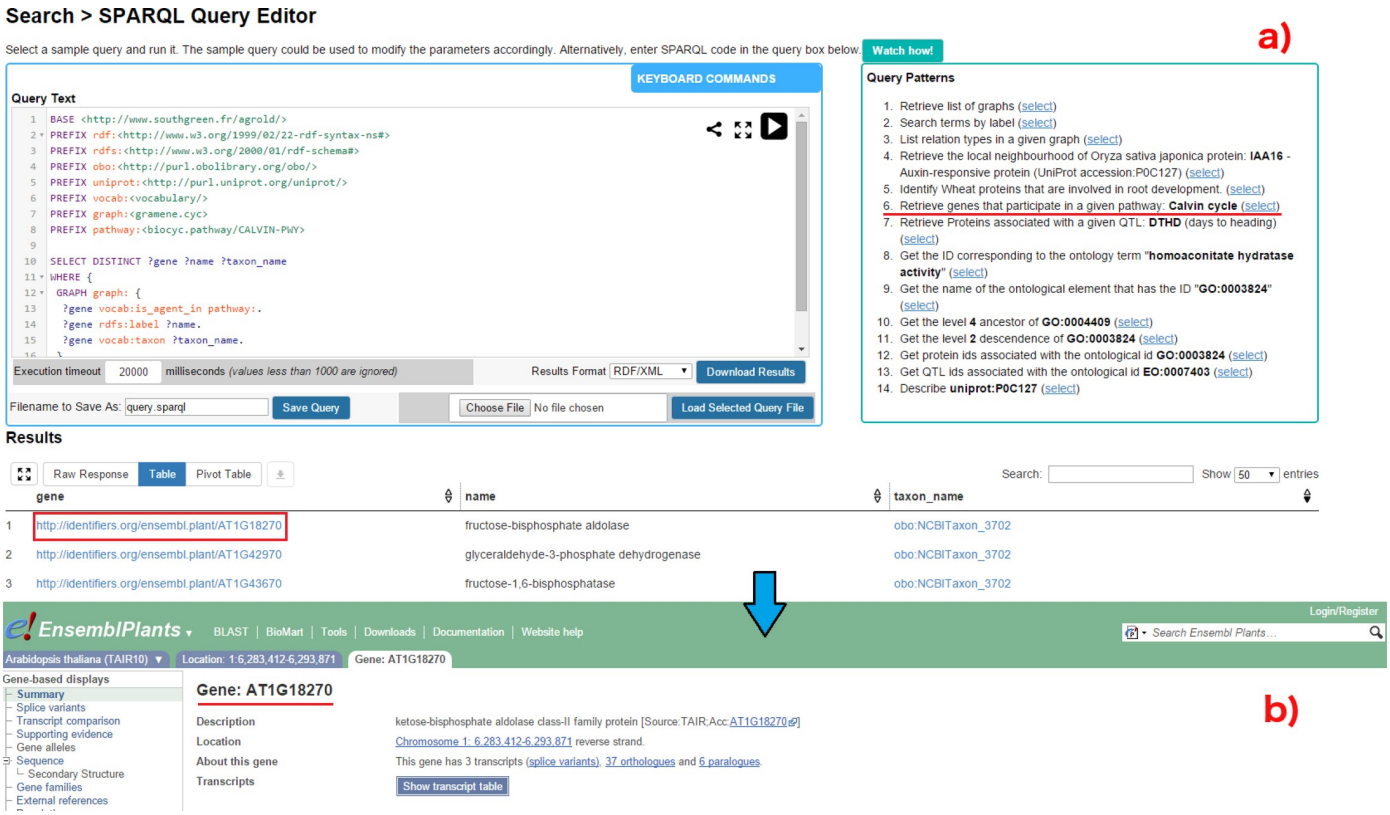


Fig 4. SPARQL query editor. Figure illustrates the execution of query Q6: (a) Q6 is one the examples queries on the top-right corner (highlighted in red). On executing the query, the results are rendered below the editor; (b) the user can look up specific genes of interest by clicking on the corresponding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

https://doi.org/10.1371/journal.pone.0198270.g004

provide non-technical users with a tool to query the knowledge base while hiding the technical aspects of SPARQL query formulation. 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.

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:

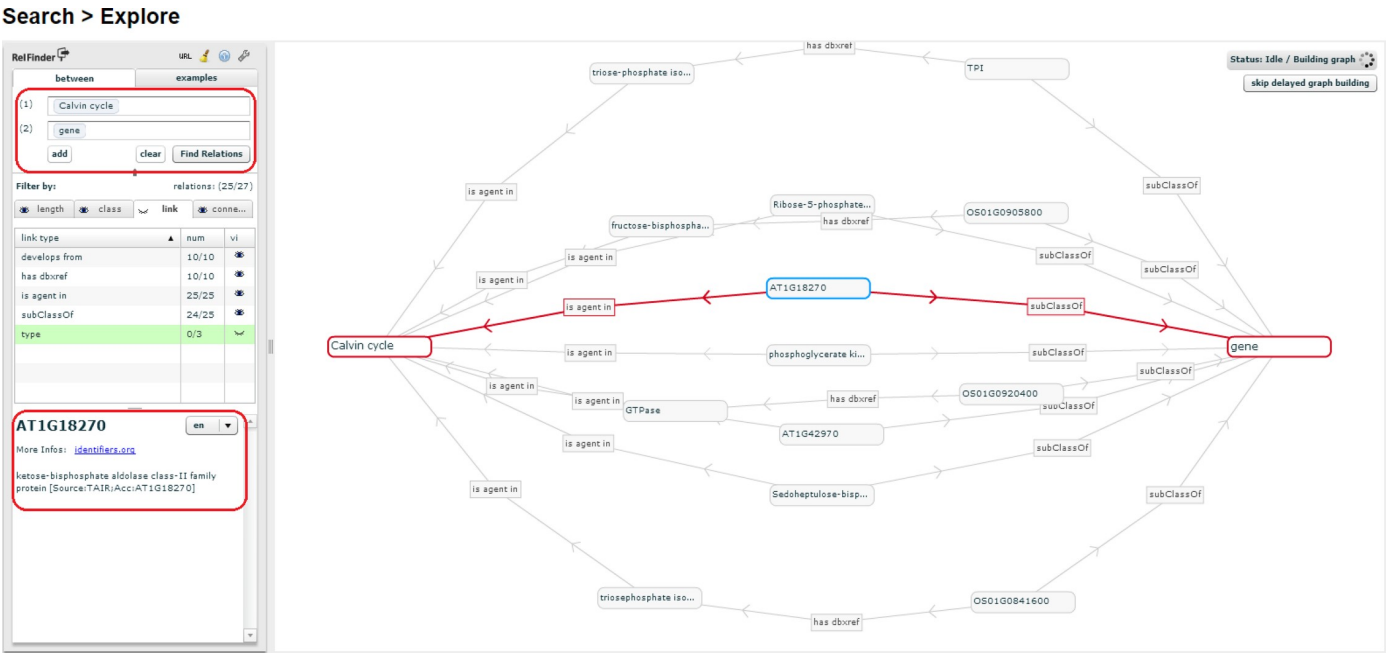


Fig 5. Exploring entity relationships in AgroLD. Figure illustrates differently the results obtained for Q6 using Explore Relationships tool. The results of Q6 can be visualized by entering the concepts (Calvin cycle and gene) in the left panel. On executing the query, all the genes involved in the chosen 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.

https://doi.org/10.1371/journal.pone.0198270.g005

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

For the evaluation of semantic search systems, Elbedweihy et al. [47] 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 [48] 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.

Recently, 20 participants were invited during 3 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). We reported the results of these sessions in S2 File as a supplementary document.

Globally, participants found the platform useful and easy to use. Overall, the idea of data navigation and traversal through knowledge graphs was well received. However, many of them needed help with some features. The general observation is that testing users ranked *Advanced*

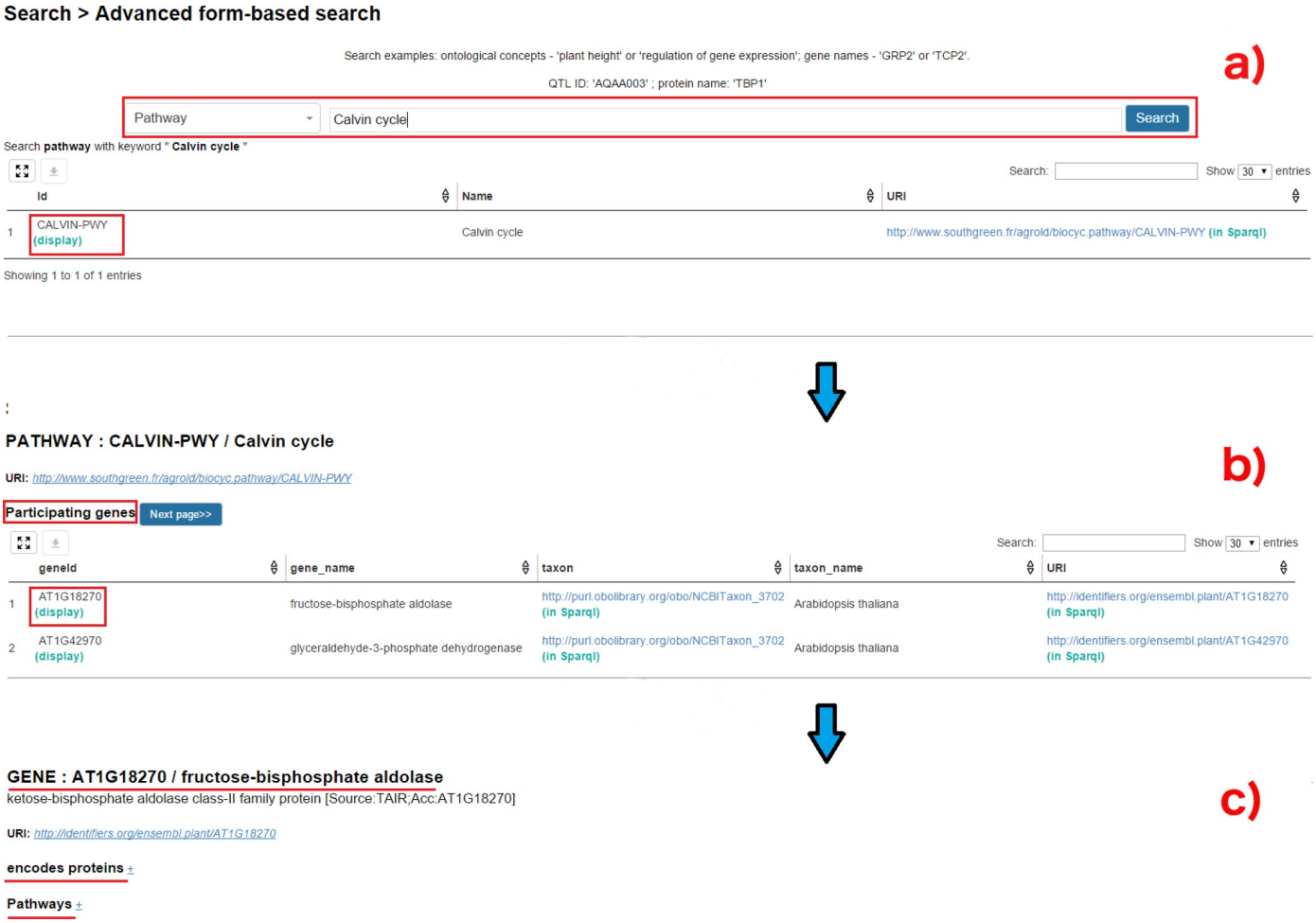


Fig 6. Advanced search query form: Figure demonstrates the steps involved in retrieving the results for Q6 using the Advanced Search query form: (a) query Q6 can be executed by selecting the type of entity (Pathways–highlighted in red) to search and entering the name of the entity (Calvin cycle). The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

https://doi.org/10.1371/journal.pone.0198270.g006

Search first then *Quick Search* after. We explain this by the display output that looks friendlier for Advanced Search. *Quick Search* won votes for usability and performance despite several comments to improve the ranking and presentation of results (4 user's comments). *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). With no surprise, evaluation results show the *SPARQL Query Editor* is the most difficult to handle. We mitigate this by offering examples of query pattern to help users handle query formulation. In the future, we will improve the examples by offering a large spectrum of search type which will follow the new phase of data integration. Furthermore, we will provide links to some SPARQL tutorials in the documentation. 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 [49] 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 and aggregates them within the result of the AgroLD query. However, this feature does not provide detailed (sentence level) assertions described in those publications. This is an area that requires further work. With the recent developments towards making text mined (sentence level) annotations available as RDF [50], query federation can be explored to retrieve entity specific assertions. This would serve as an additional provenance layer.

Limits and perspectives

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

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

These limitations identified in the current version of AgroLD will be improved in the following 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 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 [51], Gene Expression Atlas [52], on gene regulatory networks such as RiceNetDB [53] 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. 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. Additionally, we plan to fully automate the current ETL workflow.

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 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.

Supporting information

S1 File. AgroLD user guide. This document shows how to use the various features of the platform.
(PDF)

S2 File. Report of the online survey. Report of 3 sessions evaluating the AgroLD user interfaces.
(PDF)

S3 File. Examples of SPARQL queries. Example of SPARQL queries showing the benefits of property path algorithm, and complex queries.
(PDF)

S1 Table. AgroLD graph statistics.
(PDF)

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.

Author Contributions

Conceptualization: Aravind Venkatesan, Pierre Larmande.

Data curation: Aravind Venkatesan, Pierre Larmande.

Formal analysis: Aravind Venkatesan.

Funding acquisition: Manuel Ruiz, Pierre Larmande.

Investigation: Aravind Venkatesan.

Methodology: Aravind Venkatesan.

Project administration: Pierre Larmande.

Resources: Aravind Venkatesan.

Software: Aravind Venkatesan, Gildas Tagny Ngompe, Nordine El Hassouni, Imene Chentli, Valentin Guignon, Pierre Larmande.

Supervision: Pierre Larmande.

Validation: Aravind Venkatesan.

Writing – original draft: Aravind Venkatesan.

Writing – review & editing: Clement Jonquet, Manuel Ruiz, Pierre Larmande.

References

1. Goble C, Stevens R. State of the nation in data integration for bioinformatics. J Biomed Inform. Elsevier; 2008; 41: 687–693. https://doi.org/10.1016/j.jbi.2008.01.008 PMID: 18358788
2. Harper L, Campbell J, Cannon EK, Jung S, Main D, Poelchau M, et al. AgBioData Consortium Recommendations for Sustainable Genomics and Genetics Databases for Agriculture. Database. 2018; 1–7.
3. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet. 2000; 25: 25—29. https://doi.org/10.1038/75556 PMID: 10802651
4. Cooper L, Walls RL, Elser J, Gandolfo MA, Stevenson DW, Smith B, et al. The plant ontology as a tool for comparative plant anatomy and genomic analyses. Plant Cell Physiol. 2013; 54: e1. https://doi.org/10.1093/pcp/pcs163 PMID: 23220694
5. Shrestha R, Matteis L, Skofic M, Portugal A, McLaren G, Hyman G, et al. Bridging the phenotypic and genetic data useful for integrated breeding through a data annotation using the Crop Ontology

developed by the crop communities of practice. Front Physiol. 2012; 3: 326. https://doi.org/10.3389/fphys.2012.00326 PMID: 22934074

6. Buttigieg PL, Morrison N, Smith B, Mungall CJ, Lewis SE, ENVO Consortium. The environment ontology: contextualising biological and biomedical entities. J Biomed Semantics. 2013; 4: 43. https://doi.org/10.1186/2041-1480-4-43 PMID: 24330602
7. Walls RL, Deck J, Guralnick R, Baskauf S, Beaman R, Blum S, et al. Semantics in support of biodiversity knowledge discovery: an introduction to the biological collections ontology and related ontologies. PLoS One. 2014; 9: e89606. https://doi.org/10.1371/journal.pone.0089606 PMID: 24595056
8. Oellrich A, Walls RL, Cannon EK, Cannon SB, Cooper L, Gardiner J, et al. An ontology approach to comparative phenomics in plants. Plant Methods. 2015; 11: 10. https://doi.org/10.1186/s13007-015-0053-y PMID: 25774204
9. Wang Y, Wang Y, Wang J, Yuan Y, Zhang Z. An ontology-based approach to integration of hilly citrus production knowledge. Comput Electron Agric. Elsevier; 2015; 113: 24–43. https://doi.org/10.1016/J.COMPAG.2015.01.009
10. Lousteau-Cazalet C, Barakat A, Belaud J-P, Buche P, Busset G, Charnomordic B, et al. A decision support system for eco-efficient biorefinery process comparison using a semantic approach. Comput Electron Agric. Elsevier; 2016; 127: 351–367. https://doi.org/10.1016/J.COMPAG.2016.06.020
11. Jonquet C, Toulet A, Arnaud E, Aubin S, Dzalé Yeumo E, Emonet V, et al. AgroPortal: A vocabulary and ontology repository for agronomy. Comput Electron Agric. 2018; 144: 126–143. https://doi.org/10.1016/j.compag.2017.10.012
12. Berners-lee T, Hendler J, Lassila O. The Semantic Web. Sci Am. 2001; 284: 35–43.
13. W3C. Resource Description Framework (RDF): Concepts and Abstract Syntax [Internet]. [cited 3 Apr 2010]. Available: http://www.w3.org/TR/2004/REC-rdf-concepts-20040210/
14. W3C. RDF Schema 1.1 [Internet]. [cited 27 Apr 2018]. Available: https://www.w3.org/TR/rdf-schema/
15. W3C. OWL 2 Web Ontology Language Structural Specification and Functional-Style Syntax [Internet]. [cited 3 Apr 2010]. Available: http://www.w3.org/TR/2009/REC-owl2-syntax-20091027/
16. The W3C SPARQL Working Group. SPARQL 1.1 Overview [Internet]. [cited 15 Apr 2013]. Available: http://www.w3.org/TR/sparql11-overview/
17. Luciano JS, Andersson B, Batchelor C, Bodenreider O, Clark T, Denney CK, 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. https://doi.org/10.1186/2041-1480-2-S2-S1 PMID: 21624155
18. Venkatesan A, Tripathi S, Sanz de Galdeano A, Blondé W, Lægreid A, Mironov V, et al. Finding gene regulatory network candidates using the gene expression knowledge base. BMC Bioinformatics. 2014; 15: 386. https://doi.org/10.1186/s12859-014-0386-y PMID: 25490885
19. Smith RN, Aleksic J, Butano D, Carr A, Contrino S, Hu F, et al. InterMine: a flexible data warehouse system for the integration and analysis of heterogeneous biological data. Bioinformatics. Oxford University Press; 2012; 28: 3163–5. https://doi.org/10.1093/bioinformatics/bts577 PMID: 23023984
20. 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 Biotechnol. Nature Publishing Group; 2007; 25: 1251–1255. https://doi.org/10.1038/nbt1346 PMID: 17989687
21. Noy NF, Shah NH, Whetzel PL, Dai B, Dorf M, Griffith N, et al. BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Res. 2009; 37: W170–173. https://doi.org/10.1093/nar/gkp440 PMID: 19483092
22. Belleau F, Nolin M-A, Tourigny N, Rigault P, Morissette J. Bio2RDF: towards a mashup to build bioinformatics knowledge systems. J Biomed Inform. Elsevier; 2008; 41: 706–716. https://doi.org/10.1016/j.jbi.2008.03.004 PMID: 18472304
23. Williams AJ, Harland L, Groth P, Pettifer S, Chichester C, Willighagen EL, et al. Open PHACTS: Semantic interoperability for drug discovery. Drug Discovery Today. 2012. pp. 1188–1198. https://doi.org/10.1016/j.drudis.2012.05.016 PMID: 22683805
24. Momtchev V, Peychev D, Primov T, Georgiev G. Expanding the Pathway and Interaction Knowledge in Linked Life Data. International Semantic Web Challenge. 2009.
25. Jupp S, Klein J, Schanstra J, Stevens R. Developing a kidney and urinary pathway knowledge base. J Biomed Semantics. 2011; 2 Suppl 2: S7. https://doi.org/10.1186/2041-1480-2-S2-S7 PMID: 21624162
26. 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. Bioinformatics. 2014; 1–2. https://doi.org/10.1093/bioinformatics/btt765 PMID: 24413672

27. Venkatesan A, El Hassouni N, Phillipe F, Pommier C, Quesneville H, Ruiz M, et al. Towards efficient data integration and knowledge management in the Agronomic domain. APIA'15: premiere Conference Applications Pratiques de l'Intelligence Artificielle. 2015.
28. Leonelli S, Davey RP, Arnaud E, Parry G, Bastow R. Data management and best practice for plant science. Nat Publ Gr. Macmillan Publishers Limited; 2017; 3: 1–4. https://doi.org/10.1038/nplants.2017.86 PMID: 28585570
29. 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; https://doi.org/10.1093/nar/gkx1152 PMID: 29186578
30. Monaco MK, Stein J, Naithani S, Wei S, Dharmawardhana P, Kumari S, et al. Gramene 2013: Comparative plant genomics resources. Nucleic Acids Res. 2014; 42. https://doi.org/10.1093/nar/gkt1110 PMID: 24217918
31. Magrane M, Consortium UP. UniProt Knowledgebase: A hub of integrated protein data. Database. 2011;2011. https://doi.org/10.1093/database/bar009 PMID: 21447597
32. Barrell D, Dimmer E, Huntley RP, Binns D, O'Donovan C, Apweiler R. The GOA database in 2009—An integrated Gene Ontology Annotation resource. Nucleic Acids Res. 2009; 37. https://doi.org/10.1093/nar/gkn803 PMID: 18957448
33. Hamelin C, Sempere G, Jouffe V, Ruiz M. TropGeneDB, the multi-tropical crop information system updated and extended. Nucleic Acids Res. 2013; 41. https://doi.org/10.1093/nar/gks1105 PMID: 23161680
34. Droc G, Ruiz M, Larmande P, Pereira A, Piffanelli P, Morel JB, et al. OryGenesDB: a database for rice reverse genetics. Nucleic Acids Res. 2006; 34: D736–40. https://doi.org/10.1093/nar/gkj012 PMID: 16381969
35. Larmande P, Gay C, Lorieux M, Périn C, Bouniol M, Droc G, et al. Oryza Tag Line, a phenotypic mutant database for the Génoplante rice insertion line library. Nucleic Acids Res. 2008; 36: 1022–1027. https://doi.org/10.1093/nar/gkm762 PMID: 17947330
36. Conte MG, Gaillard S, Lanau N, Rouard M, Périn C. GreenPhylDB: a database for plant comparative genomics. Nucleic Acids Res. 2008; 36: D991–998. https://doi.org/10.1093/nar/gkm934 PMID: 17986457
37. 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. https://doi.org/10.1093/nar/gkv351 PMID: 26040700
38. The Gene Ontology Consortium. Gene Annotation File (GAF) specification [Internet]. [cited 20 Mar 2018]. Available: http://geneontology.org/page/go-annotation-file-format-20
39. Sequence Ontology consortium. GFF3 Specification [Internet].
40. Gibbs RA, Belmont JW, Hardenbol P, Willis TD, Yu F, Zhang H, et al. The International HapMap Project. Nature. 2003; 426: 789–796. https://doi.org/10.1038/nature02168 PMID: 14685227
41. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011; 27: 2156–8. https://doi.org/10.1093/bioinformatics/btr330 PMID: 21653522
42. Juty N, Le Novère N, Laibe C. Identifiers.org and MIRIAM Registry: community resources to provide persistent identification. Nucleic Acids Res. 2012; 40: D580–6. https://doi.org/10.1093/nar/gkr1097 PMID: 22140103
43. Smith B, Ceusters W, Klagges B, Kohler J, Kumar A, Lomax J, et al. Relations in biomedical ontologies. Genome Biol. 2005; 6: R46. https://doi.org/10.1186/gb-2005-6-5-r46 PMID: 15892874
44. Cyganiak R (National U of I, Bizer C. Pubby—A Linked Data Frontend for SPARQL Endpoints. 2008; Available: http://wifo5-03.informatik.uni-mannheim.de/pubby/
45. Heim P, Hellmann S, Lehmann J, Lohmann S, 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. https://doi.org/10.1007/978-3-642-10543-2_21
46. Rietvelda L, Hoekstraa R. The YASGUI Family of SPARQL Clients. Semant Web J. 2015; 0: 1–10.
47. Elbedweihy K, Wrigley SN, Ciravegna F, Reinhard D, Bernstein A. Evaluating semantic search systems to identify future directions of research. The Semantic Web: ESWC 2012 Satellite Events. Springer; 2012. pp. 148–162.
48. Brooke J. SUS-A quick and dirty usability scale. Usability Eval Ind. London; 1996; 189: 4–7.
49. 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. https://doi.org/10.1038/sdata.2016.18 PMID: 26978244

50. Venkatesan A, Kim J-H, Talo F, Ide-Smith M, Gobeill J, Carter J, 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. https://doi.org/10.12688/wellcomeopenres.10210.2 PMID: 28948232

51. IC4R Project Consortium, Hao L, Zhang H, Zhang Z, Hu S, Xue Y. Information Commons for Rice (IC4R). Nucleic Acids Res. 2016; 44: D1172–D1180. https://doi.org/10.1093/nar/gkv1141 PMID: 26519466

52. Petryszak R, Keays M, Tang YA, Fonseca NA, Barrera E, Burdett T, 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. https://doi.org/10.1093/nar/gkv1045 PMID: 26481351

53. Lee T, Oh T, Yang S, Shin J, Hwang S, Kim CY, et al. RiceNet v2: An improved network prioritization server for rice genes. Nucleic Acids Res. 2015; 43: W122–W127. https://doi.org/10.1093/nar/gkv253 PMID: 25813048

Computers and Electronics in Agriculture 144 (2018) 126–143

Contents lists available at ScienceDirect

Computers and Electronics in Agriculture

journal homepage: www.elsevier.com/locate/compag

Original papers

AgroPortal: A vocabulary and ontology repository for agronomy

Clément Jonquet[a,b,f,*], Anne Toulet[a,b], Elizabeth Arnaud[c], Sophie Aubin[d], Esther Dzalé Yeumo[d], Vincent Emonet[a], John Graybeal[f], Marie-Angélique Laporte[c], Mark A. Musen[f], Valeria Pesce[g], Pierre Larmande[b,e]

[a] Laboratory of Informatics, Robotics and Microelectronics of Montpellier (LIRMM), University of Montpellier & CNRS, France
[b] Computational Biology Institute (IBC) of Montpellier, France
[c] Bioversity International, Montpellier, France
[d] INRA Versailles, France
[e] UMR DIADE, IRD Montpellier, France
[f] Center for BioMedical Informatics Research (BMIR), Stanford University, USA
[g] Global Forum on Agricultural Research (GFAR), Food and Agriculture Organization (FAO) of the United Nations, Rome, Italy

ARTICLE INFO

Keywords:
Ontologies
Controlled vocabularies
Knowledge organization systems or artifacts
Ontology repository
Metadata
Mapping
Recommendation
Semantic annotation
Agronomy
Food
Plant sciences
Biodiversity

ABSTRACT

Many vocabularies and ontologies are produced to represent and annotate agronomic data. However, those ontologies are spread out, in different formats, of different size, with different structures and from overlapping domains. Therefore, there is need for a common platform to receive and host them, align them, and enabling their use in agro-informatics applications. By reusing the National Center for Biomedical Ontologies (NCBO) BioPortal technology, we have designed AgroPortal, an ontology repository for the agronomy domain. The AgroPortal project re-uses the biomedical domain's semantic tools and insights to serve agronomy, but also food, plant, and biodiversity sciences. We offer a portal that features ontology hosting, search, versioning, visualization, comment, and recommendation; enables semantic annotation; stores and exploits ontology alignments; and enables interoperation with the semantic web. The AgroPortal specifically satisfies requirements of the agronomy community in terms of ontology formats (e.g., SKOS vocabularies and trait dictionaries) and supported features (offering detailed metadata and advanced annotation capabilities). In this paper, we present our platform's content and features, including the additions to the original technology, as well as preliminary outputs of five driving agronomic use cases that participated in the design and orientation of the project to anchor it in the community. By building on the experience and existing technology acquired from the biomedical domain, we can present in AgroPortal a robust and feature-rich repository of great value for the agronomic domain.

1. Introduction

Agronomy, food, plant sciences, and biodiversity are complementary scientific disciplines that benefit from integrating the data they generate into meaningful information and interoperable knowledge. Undeniably, data integration and semantic interoperability enable new scientific discoveries through merging diverse datasets (Goble and Stevens, 2008). A key aspect in addressing semantic interoperability is the use of ontologies as a common and shared means to describe data, make them interoperable, and annotate them to build structured and formalized knowledge. Biomedicine has always been a leading domain encouraging semantic interoperability (Rubin et al., 2008). The domain has seen success stories such as the Gene Ontology (Ashburner et al., 2000), widely used to annotate genes and their products. And other disciplines have followed, developing among others the Plant Ontology (Cooper et al., 2012), Crop Ontology (Shrestha et al., 2010), Environment Ontology (Buttigieg et al., 2013), and more recently, the Agronomy Ontology (Devare et al., 2016), TOP Thesaurus (Garnier et al., 2017), Food Ontology (Griffiths et al., 2016), the IC-FOODS initiative's ontologies (Musker et al., 2016), and the animal traits ontology (Hughes et al., 2014). Ontologies have opened the space to various types of semantic applications (Meng, 2012; Walls et al., 2014), to data integration (Wang et al., 2015), and to decision support (Lousteau-Cazalet et al., 2016). Semantic interoperability has been identified as a key issue for agronomy, and the use of ontologies declared a way to address it (Lehmanna et al., 2012).

Communities engaged in agronomic research often need to access specific sets of ontologies for data annotation and integration. For instance, plant genomics produces a large quantity of data (annotated genomes), and ontologies are used to build databases to facilitate cross-

* Corresponding author at: 161 Rue Ada, 34090 Montpellier, France.
E-mail address: jonquet@lirmm.fr (C. Jonquet).

https://doi.org/10.1016/j.compag.2017.10.012
Received 27 October 2016; Received in revised form 12 May 2017; Accepted 13 October 2017

species comparisons (Jaiswal, 2011). More recently, the focus of many scientific challenges in plant breeding has switched from genetics to phenotyping, and standard traits/phenotypes vocabularies have become necessary to facilitate breeders' data integration and comparison. In parallel with very specific crop dictionaries (Shrestha et al., 2010), important organizations have produced large reference vocabularies such as Agrovoc (Food and Agriculture Organization) (Sachit Rajbhandari, 2012), the NAL Thesaurus (National Agricultural Library), and the CAB Thesaurus (Centre for Agricultural Bioscience International).[1] These thesauri are primarily used to index information resources and databases. As more vocabularies and ontologies[2] are produced in the domain, the greater the need to discover them, evaluate them, and manage their alignments (d'Aquin and Noy, 2012).

However, while great efforts have taken place in the biomedical domain to harmonize content (e.g., the *Unified Medical Language System* (UMLS), mostly for medical terminologies) (Bodenreider, 2004) and ontology design principles (e.g., the OBO Foundry, containing mostly biological and biomedical ontologies) (Smith et al., 2007), ontologies in agriculture are spread out around the web (or even unshared), in many different formats and artifact types, and with different structures. Agronomy (and its related domains such as food, plant sciences, and biodiversity) needs an one-stop shop, allowing users to identify and select ontologies for specific tasks, as well as offering generic services to exploit them in search, annotation or other scientific data management processes. The need is also for a community-oriented platform that will enable ontology developers and users to meet and discuss their respective opinions and wishes. This need was clearly expressed by stakeholders in various roles (developers, database maintainers, and researchers) across many community meetings, such as: 1st International Workshop for Semantics for Biodiversity in 2013 (http://semantic-biodiversity.mpl.ird.fr) (Larmande et al., 2013); the "Improving Semantics in Agriculture" workshop in 2015 (Baker et al., 2015); or several meetings of the Agricultural Data Interest Group (IGAD) of the Research Data Alliance.

These motivations prompted us to build a vocabulary and ontology repository to address these needs. In this paper, we present the AgroPortal project, a community effort started by the Montpellier scientific community to build an ontology repository for the agronomy domain. Our goal is to facilitate the adoption of metadata and semantics to facilitate open science in agronomy. By enabling straightforward use of agronomical ontologies, we let data managers and researchers focus on their tasks, without requiring them to deal with the complex engineering work needed for ontology management. AgroPortal offers a robust and reliable service to the community that provides ontology hosting, search, versioning, visualization, comment, and recommendation; enables semantic annotation; stores and exploits ontology alignments; and enables interoperation with the semantic web. Our vision is to facilitate the integrated use of all vocabularies and ontologies related to agriculture, regardless of their source, format, or content type.

In order to capitalize on what is already available in other communities, we have reused the openly available NCBO BioPortal technology (http://bioportal.bioontology.org) (Noy et al., 2009; Whetzel et al., 2011) to build our ontology repository and services platform. BioPortal was originally dedicated to health, biology and medicine and has some content related to agriculture, but the portal does covers few of the facets of agronomy, food, plant sciences and biodiversity, let alone environment and animal sciences. Therefore, many in the agronomy community do not see themselves as users targeted by BioPortal. For instance, the Crop Ontology is listed on the NCBO BioPortal (along with other top-level plant-related ontologies), but is not currently fully accessible and described through this portal; none of the crop specific ontologies are available. In addition to its core repository of ontology mission, the NCBO technology also offers many applicable tools, including a mapping repository, an annotator, an ontology recommender, community support features, and an index of annotated data. All these services are reused and customized within AgroPortal to benefit its target user community.[3] Furthermore, our vision was to adopt, as the NCBO did, an open and generic approach where users can easily participate to the platform, upload content, and comment on others' content (ontologies, concepts, mappings, and projects). As explained below, we determined that the NCBO technology (Whetzel and Team, 2013) implemented the greatest number of our required features, while recognizing the technical challenges of adopting such a various and complex software.

In the following sections, we offer extensive descriptions of AgroPortal's features. We will focus on how they address community requirements expressed within five agronomic driving use cases involving important research organizations in agriculture such as Bioversity International (CGIAR), French INRA, and United Nations FAO. The rest of the paper is organized as follows: In Section 2, we review related work in ontology repositories in relation to our domain of interest. Section 3 describes the requirements of AgroPortal's initial five driving agronomic use cases. Section 4 presents our platform by extensively describing its content, as well as its features (both inherited from the NCBO BioPortal, and added by us). Section 5 analyzes how our initial five driving use case results benefit from AgroPortal. Finally, Section 6 provides a discussion of the contributions of AgroPortal, and Section 7 presents our conclusions.

2. Background and related work

With the growing number of developed ontologies, ontology libraries and repositories have been of interest in the semantic web community. Ding and Fensel (2001) presented in 2001 a review of ontology libraries that introduced the notion of "library." Then Hartman et al. Baclawski and Schneider (2009) introduced the concept of ontology repository, with advanced features such as search, metadata management, visualization, personalization, and mappings. By the end of the 2000's, the Open Ontology Repository Initiative (Baclawski and Schneider, 2009) was a collaborative effort to develop a federated infrastructure of ontology repositories.[4] d'Aquin and Noy (2012) provided the latest review of ontology repositories in 2012.

In the biomedical or agronomic domains there are several standards or knowledge organization systems libraries (or registries) such as FAIRSharing (http://fairsharing.org) Sansone et al., 2012, the FAO's VEST Registry (http://aims.fao.org/vest-registry), and the agINFRA linked data vocabularies (vocabularies.aginfra.eu) (Pesce et al., 2013). They usually register ontologies and provide a few metadata attributes about them. However, because they are registries not focused on vocabularies and ontologies, they do not support the level of features that an ontology repository offers. In the biomedical domain, the OBO Foundry (Smith et al., 2007) is a reference community effort to help the

[1] http://aims.fao.org/agrovoc, https://agclass.nal.usda.gov and http://www.cabi.org/cabthesaurus

[2] In this paper, we often use the word "ontologies" or "vocabularies and ontologies" to include ontologies, vocabularies, terminologies, taxonomies and dictionaries. We acknowledge the differences (not discussed here) in all these types of Knowledge Organization Systems (KOS) or knowledge artifacts. The reader may refer to McGuinness's discussion (McGuinness, 2003). While being an "ontology repository", AgroPortal handles all these artifact types, if they are compatibly formatted. While AgroPortal thereby enables horizontal use of these artifact types with common user interface and application programming interface, it does not leverage the full power of ontologies (e.g., reasoning), instead map all the imported artifact types to a "common simplified model."

[3] Except the "NCBO Resource Index" component, a database of 50+ biomedical resources indexed with ontology concepts (Jonquet et al., 2011) that we have not reused in AgroPortal because we work with the AgroLD use case to fulfill the mission of interconnecting ontologies and data.

[4] At that time, the effort already reused the NCBO technology that was open source, but not yet packaged in an appliance as it is today.

biomedical and biological communities build their ontologies with an enforcement of design and reuse principles that have made the effort very successful. The OBO Foundry web application is not an ontology repository per se, but relies on other applications that pull their data from the foundry, such as the NCBO BioPortal (Noy et al., 2009), OntoBee (Xiang et al., 2011), the EBI Ontology Lookup Service (Côté et al., 2006) and more recently AberOWL (Hoehndorf et al., 2015). In addition, there exist other ontology libraries and repository efforts unrelated to biomedicine, such as the Linked Open Vocabularies (Vandenbussche et al., 2014), OntoHub (Till et al., 2014), and the Marine Metadata Initiative's Ontology Registry and Repository (Graybeal et al., 2012).

Some of the known ontology repositories could be candidates for hosting agronomical ontologies. However, all of these portals either are too generic, or too narrowly focused on health, biology or medicine, and despite any existing thematic overlaps, scientific lineage and partnerships, we have identified, as established in Section 1, the crucial need for a community platform where agronomy will actually be the primary focus. To avoid building a new ontology repository from scratch, we have considered which of the previous technologies are reusable. While all of them are open source, only the NCBO BioPortal[5] and OLS[6] are really meant for reuse, both in their construction, and in their provided documentation. At the start of our project in 2014, AberOWL was not yet published and OntoBee (released in 2011) had not changed between 2011 and 2014 (a new release took place thereafter (Ong et al., 2016). Of the two candidate technologies at the time, we will show, that the NCBO technology was the one implementing highest number of requested features.[7]

In the biomedical domain, the NCBO BioPortal is a well-known open repository for biomedical ontologies originally spread out over the web and in different formats. There are 656 public ontologies in this collection as of Nov. 2017, including relevant ones for agronomy. By using the portal's features, users can browse, search, visualize and comment on ontologies both interactively through a user web interface, and programmatically via web services. Within BioPortal, ontologies are used to develop an annotation workflow (Jonquet et al., 2009) that indexes several biomedical text and data resources using the knowledge formalized in ontologies to provide semantic search features that enhance information retrieval experience (Jonquet et al., 2011). The NCBO BioPortal functionalities have been progressively extended in the last 12 years, and the platform has adopted semantic web technologies (e.g., ontologies, mappings, metadata, notes, and projects are stored in an RDF[8] triple store) (Salvadores et al., 2013).

An important aspect is that NCBO technology (Whetzel and Team, 2013) is domain-independent and open source. A BioPortal virtual appliance[9] is available as a server machine embedding the complete code and deployment environment, allowing anyone to set up a local ontology repository and customize it. It is important to note that the NCBO Virtual Appliance has been quite regularly reused by organizations which needed to use services like the NCBO Annotator but, for privacy reason, had to process the data in house. Via the Virtual Appliance, NCBO technology has already been adopted for different ontology repositories in related domains and was also chosen as foundational software of the Open Ontology Repository Initiative (Baclawski and Schneider, 2009). The Marine Metadata Interoperability Ontology Registry and Repository (Rueda et al., 2009) used it as its backend storage system for over 10 years, and the Earth Sciences Information Partnership earth and environmental semantic portal (Pouchard and Huhns, 2012) was deployed several years ago. More recently, the SIFR BioPortal (Jonquet et al., 2016) prototype was created at University of Montpellier to build a French Annotator and experiment multilingual issues in BioPortal (Jonquet et al., 2015). Although we cannot know all the applications of other technologies, the visibly frequent reuse of the NCBO technology definitively confirmed it was our best candidate. There are two other major motivations for AgroPortal to reuse the products of biomedicine: (i) to avoid re-developing tools that have already been designed and extensively used and contribute to long term support of the commonly used technology; and (ii) to offer the same tools, services and formats to both communities, to facilitate the interface and interaction between their domains. This alignment will enhance both technical reuse (for example, enabling queries to either system with the same code), and semantic reuse (knowing the same semantic capabilities and practices apply to both sets of ontologies).

More specifically to the plant domain, the Crop Ontology web application (www.cropontology.org) (Matteis et al., 2013) publishes online sets of ontologies and dictionaries required for describing crop germplasm, traits and evaluation trials. As of Nov. 2017, it contains 28 crop-specific phenotype and trait ontologies, in addition to ontologies related to the crop germplasm domain. Besides its role as a repository, the Crop Ontology web application offers community-oriented features such as an CSV template (TDv5) for trait submission, and addition and filtering of new terms. A web Application Programming Interface (API) provides all necessary services to third party users like the Global Evaluation Trials Database, currently storing 35,000 trial records. Efforts have been made to structure and formalize the crop-specific ontologies following semantic web standards (using the Web Ontology Language (OWL)), as well as offering collaborative ontology enrichment and annotation features. The current Crop Ontology web application facilitates the ontology-engineering life cycle (Noy et al., 2010), starting with collaborative construction, publishing, use and modification. However, it would require important improvements such as: versioning, community features, multilingual aspects, visualization, data annotation, and mapping services. For instance, it is important to support the alignment (or mapping) of terms within and across different ontologies both within the Crop Ontology itself (in different crop branch) and with other top level ontologies commonly used in plant biology, like the Plant Ontology, Plant Trait Ontology, Plant Environment Ontology, Plant Stress Ontology all maintained and extended within the Planteome project (Jaiswal et al., 2016).

The Planteome platform (www.planteome.org) is reusing the Gene Ontology project AmiGO technology (Carbon et al., 2009) to build a database of searchable and browsable annotations for plant traits, phenotypes, diseases, genomes, gene expression data across a wide range of plant species. The project focuses on developing reference ontologies for plant and on integrating annotated data within the platform. Their objective is slightly different than AgroPortal's objective, and the scope is not as large as the one we envision for AgroPortal.

3. Driving agronomic use cases requirements

The AgroPortal project was originally driven by five agronomic use cases that were the principal sources of ontologies and vocabularies. In this section, we present their requirements in terms of ontology repository functionalities – summarized in Table 1. The results for each use case will be presented in Section 5.

[5] The technology has always been open source, and the appliance has been made available since 2011. However, the product became concretely and easily reusable after BioPortal v4.0 end of 2013.

[6] The technology has always been open source but some significant changes (e.g., the parsing of OWL) facilitating the reuse of the technology for other portals were done with OLS 3.0 released in December 2015.

[7] It is beyond the scope of this paper to draw a complete comparison of ontology portals. The reader may refer to d'Aquin and Noy (2012).

[8] The Resource Description Framework (RDF) is the W3C language to described data. It is the backbone of the semantic web. SPARQL is the corresponding query language. By adopting RDF as the underlying format, AgroPortal can easily make its data available as linked open data and queryable through a public SPARQL endpoint. To illustrate this, the reader may consult the Link Open Data cloud diagram (http://lod-cloud.net) that since 2017 includes ontologies imported from the NCBO BioPortal (most of the Life Sciences section).

[9] www.bioontology.org/wiki/index.php/Category:NCBO_Virtual_Appliance

Table 1
Summary of agronomic use case requirements for AgroPortal.

#	Requirement	Use case	Example
1	One-stop-shop to store, browse, search, visualize agronomical ontologies	LovInra	Facilitate the adoption of semantic web standards by INRA' scientist, with a focus on agriculture
		VEST	The registry targets specifically the agriculture community and requires content-based services. The organization of ontologies by group and categories is also necessary
2	Unique ontology access point and application programming interface (API) to ontologies	AgroLD	Automatically retrieve the most recent version of ontologies currently hosted either on OBO Foundry or Cropontology.org. At the beginning of the project, a SPARQL endpoint for ontologies was also needed
		VEST	Access point to automatically obtain metadata about all the ontologies
3	Directly accessible to scientists to upload their ontologies or vocabularies	LovInra, VEST	INRA's researchers and VEST users need to upload their resources to a platform themselves
4	Ontology-based annotation service	AgroLD	Annotate text data from database fields to create RDF triples
		LovInra, Crop Ontology	Identify plant phenotypes in text descriptions
5	Handle different level of semantic description and the corresponding standard formats (SKOS and OWL)	LovInra	INRA's develop different type of knowledge organization systems include: ontologies (AFEO, Biorefinery, OntoBiotope) but also thesauri (AnAEE, GACS)
		VEST	Many resources in agronomy are in SKOS format.
6	Store and retrieve mappings between ontologies	ALL	All use cases have expressed the need to have a place to store, describe and retrieve alignments
7	Store mappings between ontologies and external resources	AgroLD	Publish AgroLD mapping annotations to reference ontologies such as SIO, EDAM, PO
		Others	Reference thesauri like Agrovoc have adopted linked open data practices and offer mappings to multiple semantic web resources (not necessarily ontologies)
8	Automatically generate mappings between ontologies	ALL	All use cases have expressed the need to automatically align ontologies one another
9	Query and search annotated data from ontologies	AgroLD	Identify AgroLD data elements when browsing ontologies in AgroPortal.
10	Offer a unique sub-endpoint specific to a community or group	WDI	Visualize and use only the 22 vocabularies identified by the WDI working group
		LovInra	Clearly identify resources (co-)developed by INRA's researchers
		Crop Ontology	Handle as a collection the Crop Ontology project, which is composed of multiple crop-specific trait ontologies. Possible alternative to cropontology.org
11	Provide rich metadata description for ontologies (using semantic web standards)	WDI	Clearly describe access rights and license information for ontologies
		LovInra	Clearly describe the type of resources (ontology, thesaurus, vocabulary, etc.) and their format and syntax
		VEST	Facilitate an automatic interconnection with VEST, including aligning the metadata fields
12	Get community feedback	WDI	Inform the community about the WDI guidelines and get their feedback on the selected ontologies
		Crop Ontology	Offer breeders a way to suggest new trait and comment existing ones
		VEST	Enable a large community of "standard" developers to provide feedback and comments on the use (or non-use) of ontologies and vocabularies in AgroPortal
13	Multilingual ontology support	VEST, Others	Increasingly vocabularies have labels in different languages (e.g., Agrovoc, GACS, NALt). Distinguish between these labels in lexical-based services (search, annotation)
		Others	IRSTEA develops vocabularies only in French
14	Dereference URIs for ontologies	LovInra, Crop Ontology	When opening in a web browser a URI created by INRA or CO, display the corresponding class or property page
15	Mechanism to identify and select the relevant ontologies for a given task	LovInra, VEST	Facilitate the identification of relevant agronomical ontologies for non-experts
16	Enable private access to ontologies during working and/or development phases	LovInra	Access and test the AnAEE Thesaurus or GCAS before they release; work on certain versions of OntoBiotope not public in OpenMinted project
17	Export ontologies in different formats, including downgrading them to CSV	Crop Ontology	Breeders may need simpler formats, as they may not be able to use advanced semantic web formats
18	Store the project/ontology relationships	VEST, AgBioData	Select and maintain a list of ontologies used by model organism databases

3.1. Agronomic Linked Data (AgroLD)

Agronomic research aims to effectively improve crop production through sustainable methods. To this end, there is an urgent need to integrate data at different scales (e.g., genomics, proteomics and phenomics). However, available agronomical information is highly distributed and diverse. Semantic web technology offers a remedy to the fragmentation of potentially useful information on the web by improving data integration and machine interoperability (Schmachtenberg et al., 2014). This has been often illustrated in data integration and knowledge management in the biomedical domain (Belleau et al., 2008; Jonquet et al., 2011; Jupp et al., 2014; Groth et al., 2014). To further build on this line of research in agronomy, we have developed the Agronomic Linked Data knowledge base (www.agrold.org) (Venkatesan et al., 2015). Launched in May 2015, it serves as a platform to consolidate distributed information and facilitate formulation of research hypotheses. AgroLD offers information on genes, proteins, Gene Ontology Associations, homology predictions, metabolic pathways, plant traits, and germplasm, on the following species: rice, wheat, arabidopsis, sorghum and maize. We provide integrated agronomic data, as well as the infrastructure to aid domain experts answering relevant biological questions (for example, "identify wheat proteins that are involved in root development"). AgroLD relies on RDF and SPARQL technologies for information modelling and retrieval, and uses OpenLink Virtuoso (version 7.1) triple store. Database contents were parsed and converted into RDF using a semi-automated pipeline implemented in Python (https://github.com/SouthGreenPlatform/AgroLD).

The conceptual framework for knowledge in AgroLD is based on well-established ontologies in plant sciences such as Gene Ontology, Sequence Ontology, Plant Ontology, Crop Ontology and Plant Environment Ontology. AgroLD needs a dedicated application programming interface to these ontologies, as well as a means to annotate database fields (header and values) with ontology concepts. In addition, it requires a system to store mappings annotations between key entities in the AgroLD knowledge base and reference ontologies. In the long-term vision for AgroPortal and AgroLD, the former might be an entry point to the knowledge stored in AgroLD, enabling users to easily query and locate data annotated with ontologies.

3.2. RDA Wheat Data Interoperability (WDI) working group

Wheat is a major source of calories and protein, especially for consumers in developing countries, and thus plays an important socio-economical role. The International Wheat Initiative (www.wheatinitiative.org) has identified easy access and interoperability of all wheat related data as a top priority, to make the best possible use of genetic, genomic and phenotypic data in fundamental and applied wheat science. For example, the identification of causative genes for an important agronomic trait is key to effective marker-assisted breeding and reverse genetics. It requires integrating information from many different sources such as gene function annotations, biochemical pathways, gene expression data, as well as comparative information from related organisms, gene knock-out and the scientific literature (Hassani-Pak et al., 2013). However, the disparate nature of the formats and vocabularies used to represent and describe the data has resulted in a lack of interoperability.

The Wheat Data Interoperability working group was created in March 2014 within the frame of the Research Data Alliance (https://rd-alliance.org) and under the umbrella of the International Wheat Initiative, in order to provide a common framework for describing, linking and publishing wheat data with respect to existing open standards. The working group conducted a survey to identify and describe the most relevant vocabularies and ontologies for data description and annotation in the wheat domain (Dzalé-Yeumo et al., 2017). For some data types like DNA sequence variations, genome annotations, and gene expressions, the survey showed good consensus regarding data exchange formats. However, the survey did not show good consensus about data exchange formats and data description practices for phenotypes and germplasm, suggesting the need for harmonization and standardization.

Finally, this group identified 22 relevant vocabularies and ontologies for which, beyond the consensus issue, other problems were identified: (i) format and location heterogeneity: ontology formats included OBO format, OWL, and even SKOS (or SKOS-XL); (ii) heterogeneity: these ontology coverages ranged from describing generic experimental crop study (e.g., Crop Research), to narrow wheat-related topics (Wheat Trait, Wheat Anatomy and Development), to top-level concepts in biomedicine (BioTop). The need to offer a dedicated repository of linked vocabularies and ontologies relevant for wheat having been identified, the NCBO technology was seen as a likely tool to address this needs and desired features.

3.3. INRA Linked Open Vocabularies (LovInra)

What does a specialist in cattle developmental biology really need to easily identify, evaluate and exploit a few potential vocabularies of interest? Whether familiar with semantics technology or not, she needs a place that reflects her scientific environment and community, where those with similar concerns can share comments and content. As an example, INRA develops models to predict feed efficiency and meat quality for beef production, using experimental data collected during decades at INRA and externally. To meet the challenge of data integration, INRA developed the Animal Trait Ontology for Livestock (ATOL). In part thanks to AgroPortal, ATOL developers have identified the Animal Disease Ontology (ADO), developed by another team at INRA, as a possible resource to expand the perimeter of actionable data. This raised the question: How many complementary or competing resources to ATOL exist?

With this vision in mind, LovInra is a service offered by the French National Institute for Agricultural Research (INRA) Scientific and Technical Information department to identify and evaluate knowledge organization sources produced by INRA's scientists, so that the agricultural community and possibly a larger public can benefit from them. Many such resources developed within specific projects remain unknown to the research community despite their value. They are often developed by subject matter experts who are not semantic experts, and who often do not have the resources (knowledge, time, or money) to share their results. Further, they span multiple semantic levels, from simple lexical descriptions, to hierarchies, to complex semantic relations. To achieve this goal, the vocabularies must be published with respect to open standards and linked to other existing resources. INRA adopted the semantic web's practices and standards (RDF, SKOS, OWL, SPARQL) to enable the methodological and technical practices needed by INRA's scientists to standardize, document and publish the vocabularies created in their projects. Examples of INRA's projects developing vocabularies or ontologies includes: (i) the AnAAE Thesaurus for the semantic description of the study of continental ecosystems developed by the AnaEE-France infrastructure;[10] (ii) the OntoBiotope ontology of microorganism habitats used collaboratively in multiple projects such as OpenMinted as well as for the BioNLP shared tasks; (iii) the Agri-Food Experiment Ontology (AFEO) ontology network which cover various viticultural practices, and winemaking products and operations.

Beyond its evaluation and standardization role, LovInra also serves to assign, deference, and provide programmatic access to INRA URIs (for example, http://opendata.inra.fr/ms2o/Observation), using its triple store and web interface (http://lovinra.inra.fr). Although the current service, which includes description of resource metadata and direct access to source files, is necessary for internal use, it does not meet external dissemination objectives. In addition, the LovInra registry does not support any content-based features, such as searching, browsing, visualizing, mappings and annotation. We see AgroPortal as a possible solution to the entire range of INRA's unmet semantic needs above, complementing the services already provided by LovInra.

3.4. The Crop Ontology project

Communities engaged in germplasm evaluation trials need to access specific sets of ontologies for plant data annotation and integration. The Crop Ontology project (www.cropontology.org) (Shrestha et al., 2010) of the Integrated Breeding Platform (IBP) is AgroPortal's fourth use case. The main goals of this project are: (i) to publish online fully documented lists of breeding traits and standard variables used for producing standard field books and (ii) to support data analysis and integration of genetic and phenotypic data through harmonized breeders' data annotation (Shrestha et al., 2012). Crop breeders, data managers, modelers, and computer scientists created a community of practice to discuss their variables, methods and scales of measurement, and field books. They seek to develop the most complete crop-specific trait ontologies according to the Crop Ontology template and guidelines.

The Crop Ontology website, released in 2010, provides 28 crop-specific trait ontologies, in addition to ontologies describing germplasm material and evaluation trials. The website publishes each crop-specific trait ontology online, making it available for download from the user interface or through an API in various formats: CSV, OBO, RDF/SKOS. Partners like the Oat Global, the US Department of Agriculture (USDA), INRA and the Polish Genomic Network have uploaded ontologies.[11] The project requires a specific dedicated infrastructure that deals with the adopted multi-trait ontologies approach, and supports search and versioning of ontologies. Plus, the Crop Ontology breeders need an interface to suggest new crop traits (i.e., new terms in the trait ontologies) and simple formats (such as CSV) to export the "trait dictionary" locally.

[10] Analysis and Experimentation on Ecosystems is European research infrastructure dedicated to the experimental manipulation of managed and unmanaged terrestrial and aquatic ecosystems (www.anaee.com).

[11] In addition, the Crop Ontology is used by several third-party projects like the Next Generation Breeding (Nextgen) databases, the Integrated Breeding Platform's breeding management system, and the global repository of the Agricultural Trials or EU-SOL.

3.5. GODAN Map of Agri-Food Data Standards

Recently, a new project under the umbrella of the GODAN[12] initiative called *GODAN Action* identified as one of its outputs a global map of standards used for exchanging data in the field of food and agriculture. To avoid duplicating effort, and to reuse previous community work, the project reviewed possible sources of standards that could be integrated. Two existing suitable platforms were identified: the FAO Agricultural Information Management Standards VEST Registry (http://aims.fao.org/vest-registry – now merged inside the new Map of Standards presented Section 5.5) and the then-new AgroPortal project.

The VEST Registry, created by FAO in 2011, was a metadata catalog of around 200 knowledge organization sources and tools. It had a broader coverage than the AgroPortal in two facets, knowledge types and domains. (i) Types of vocabularies or standards covered: the VEST Registry covered all types of knowledge artifacts, not just vocabularies or ontologies formally defined in RDFS, OWL, SKOS, or OBO. For instance, the VEST registry would cover data exchange format specification defined in XML or text description. (ii) Domain coverage: Besides standards used specifically for food and agriculture data, the directory included resources used in neighboring disciplines (like climate and environment, sciences). The VEST Registry was conceived as a metadata catalog, providing descriptions and categorization of standards and linking to the original website or download of the standard, but it did not exploit the content of the vocabularies or ontologies, only their metadata descriptions. It did not support any alignment between the sources either. To interconnect the VEST and AgroPortal, rich and unambiguous metadata would be crucial, as well as good classification of resources per categories and types.

3.6. Other requirements identified

In addition to these five first driving use cases, other projects or organizations have identified AgroPortal as a relevant application to host, share and serve their ontologies:

> IRSTEA's projects, such as the French Crop Usage thesaurus about crops cultivated in France, and the French Agroecology Knowledge Management ontology for design innovative crop systems. These two projects produce ontologies only in French and needed a host for their work.
>
> The Agrovoc thesaurus (Sachit Rajbhandari, 2012), which is the most worldwide used multilingual vocabulary developed by FAO. Agrovoc contains more than 32 K concepts covering topics related to food, nutrition, agriculture, fisheries, forestry, environment and other related domains. Agrovoc Linked Open Data version contains multiple mappings to other vocabularies or resources that a resource hosting Agrovoc must incorporate.
>
> The Consortium of Agricultural Biological Databases (www.agbiodata.org), a group of database developers and curators maintaining model organism databases. The group wants to identify which databases use which ontologies, and recommend a list of ontologies based on that information.

4. A portal for agronomic related ontologies

In 2014, the *Computational Biology Institute of Montpellier* project identified the need for an ontology-based annotation service for the AgroLD and Crop Ontology use cases above. This large bioinformatics project in France had a specific plant/agronomy data work package. In parallel, we started reusing NCBO technology (Whetzel and Team, 2013) in the context of the SIFR[13] project, in which we develop a French version of the Annotator (Jonquet et al., 2016). We then implemented a connector to BioPortal within WebSmatch (an open environment for matching complex schemas from many heterogeneous data sources (Coletta et al., 2012) enabling calls either to the NCBO Annotator web service, or any other NCBO-based Annotator (Castanier et al., 2014). Once we had a portal prototype hosting a few specific ontologies, interest in it grew when we presented it to several interlocutors (for examples, Bioversity International, INRA, IRD, CIRAD, FAO, RDA, Planteome). Driven additionally by the other use cases presented in Section 3, we extended our reuse of the NCBO technology to the full stack, and publish it under the brand AgroPortal.

We now have an advanced prototype platform (illustrated in figures on following pages) whose latest version v1.4 was released in July 2017 at http://agroportal.lirmm.fr.[14] The platform currently hosts 77 ontologies (Table 2), with more than 2/3 of them not present in any similar ontology repository (like NCBO BioPortal), and 11 private ontologies. We have identified 93 other candidate ontologies (Table 3) and we work daily to import new ones while involving/informing the original ontology developers. The platform already has more than 90 registered users. For an overview of AgroPortal ontology analytics, see Fig. 5 (Annex).

4.1. Ontology organization and sources

Developers generally upload their ontologies when they think the ontologies have reached a sufficient maturity and relevance to make them publicly available. Sometime, like in the AnaEE thesaurus, or OntoBiotope, developers use/used the portal as a staging location before the ontology goes public. If the initiative comes from our side, we usually always interact with the developers before importing any new resources: the original ontology developers always stay the only authority for the ontologies in the portal. Because of the features offered by AgroPortal (Sections 4.2 and 4.3), we think it is reasonable to incorporate ontologies that are already listed on other platforms (OBO Foundry, FAIRSharing, VEST registry, or LovInra). However, in those cases we follow these practices:

> Developers can configure the entry in AgroPortal to automatically pull new version of ontologies. We synchronize the ontology in AgroPortal with the one at the original location via a nightly update[15] so the latest version is always available. For instance, all the ontologies in the OBO-FOUNDRY group are systematically updated using their PURL (e.g., for the Plant Ontology: http://purl.obolibrary.org/obo/po.owl).
>
> We always inform the ontology developers of their ontology publication on AgroPortal if they did not submit their ontology directly, and offer them to claim administration role on the ontology if desired. While we often edit ontology descriptions, we ask the ontology developers to validate our edits and complete them.
>
> We try to avoid duplicating ontologies already hosted in the NCBO BioPortal, unless required by a specific use case. Of course, overlap exists between our domain of interest and biomedicine. Our general approach is to let ontology developers decide if their ontology should be incorporated in the AgroPortal while it is already in the NCBO BioPortal. The long-term vision for AgroPortal and BioPortal is an interconnected network of "bioportals" that will enable easy access to ontologies for anyone independently from where they are hosted and that could extend to ontology repository types beyond the NCBO technology.

[12] Global Open Data for Agriculture and Nutrition: http://www.godan.info.

[13] Semantic Indexing of French Biomedical Data Resources (SIFR) project - http://www.lirmm.fr/sifr.

[14] https://github.com/agroportal/documentation/wiki/Release-notes

[15] Except for three ontologies (GO, BIOREFINERY & TRANSMAT) that are updated only weekly for scalability reasons.

Table 2
Examples of ontologies uploaded in AgroPortal. Acronyms in parenthesis are the identifier on AgroPortal e.g., http://agroportal.lirmm.fr/ontologies/AEO has the acronym AEO (Size = approximate number of classes or concepts).

Title	Format	Source	Group	Size
IBP rice trait ontology (CO_320)	OWL	cropontology.org	CROP, AGBIODATA, AGROLD	~2K
IBP wheat trait ontology (CO_321)	OWL	cropontology.org	CROP, AGBIODATA, AGROLD, WHEAT	~1K
IBP wheat anatomy & development ontology (CO_121)	OBO	cropontology.org	CROP, WHEAT	~80
IBP crop research (CO_715)	OBO	cropontology.org	CROP, AGBIODATA, WHEAT	~250
Multi-crop passport ontology (CO_020)	OBO	cropontology.org	CROP	~90
Biorefinery (BIOREFINERY)	OWL	Inra	LOVINRA, WHEAT, AGBIODATA	~300
Matter transfer(TRANSMAT)	OWL	Inra	LOVINRA, WHEAT, AGBIODATA	~1.1 K
Plant ontology (PO)	OWL	OBO Foundry	OBOF, AGROLD, WHEAT, AGBIODATA	~2K
Plant trait ontology (TO)	OWL	OBO Foundry	OBOF, AGROLD, WHEAT, AGBIODATA	~4.4 K
Durum wheat (DURUM_WHEAT)	OWL	Inra	LOVINRA	~130
Agricultural experiments (AEO)	OWL	Inra	LOVINRA	~60
Environment ontology (ENVO)	OWL	OBO Foundry	WHEAT, OBOF	~6.3 K
NCBI organismal classification (NCBITAXON)	RRF	UMLS	WHEAT, AGROLD	~900 K
AnaEE thesaurus (ANAEETHES)	SKOS	Inra	LOVINRA	~3.3 K
French crop usage (CROPUSAGE)	SKOS	Irstea	None	~300
Agrovoc (AGROVOC)	SKOS	FAO (UN)	WHEAT, AGBIODATA	~32 K
Food ontology (FOODON)	OWL	OBO Foundry	OBOF	~10 K
National agricultural library thesaurus (NALT)	SKOS	NAL (USDA)	WHEAT, AGBIODATA	~67 K
Global agricultural concept scheme (GACS)	SKOS	FAO-NAL-CABI	None	~580 K
Agronomy ontology	OWL	CGIAR	OBOF	~430
Biological collections ontology	OWL	OBO Foundry	OBOF	~160
Flora phenotype ontology	OWL	AberOWL	None	~28 K

Table 3
Selection of candidate ontologies of interest for the agronomic community, not present in the NCBO BioPortal.

Title	Organization or source
CAB thesaurus	CABI
Chinese agricultural thesaurus	CAAS
Wine ontology	INRA
Oat, Barley, Brachiaria, Potato (etc.) trait ontologies	Crop Ontology
Plant disease ontology	INRA
Agriculture activity ontology	CAVOC
Agriculture and forestry ontology	Univ. of Helsinki
IC-FOODS ontologies (~10)	UC Davis
agINFRA soil vocabulary	FAO, GFAR
Plant-pathogen interactions ontology	CBGP
Plant phenology ontology	OBO Foundry
Thesaurus of plant characteristics	CEFE
Livestock product trait ontology	Iowa State Univ.
Livestock breed ontology	Iowa State Univ.

Within AgroPortal, each time an ontology is uploaded into the portal, it is assigned a group and/or category. Groups associate ontologies from the same project or organization, for better identification of the provenance. We have created a group for each use case, except the fifth one that is not a source of ontologies, and another one for the OBO Foundry. For each group we have deployed a specific slice (a restriction of the user interface to a specific group of ontologies) as explained later. Categories indicate the topic(s) of the ontology, providing another way to classify ontologies in the portal independently from their groups or provenance. As of now we have defined 20 general categories such as Farms and Farming Systems, Plant Phenotypes and Traits, Plant Anatomy and Development, Agricultural Research, and Technology and Engineering. These categories were established in cooperation with FAO Agricultural Information Management Standards (AIMS), which has maintained the VEST Registry since 2011.

Groups and categories, along with other metadata, can be used on the "Browse" page of AgroPortal to filter out the list of ontologies (cf. Fig. 3). Of course, groups and categories are customizable, and will be adapted in the future to reflect the evolution of the portal's content and community feedback. The portal's architecture provides URIs for any portal objects, including groups and categories. For example, the URI http://data.agroportal.lirmm.fr/categories/FARMING identifies the group "Farms and Farming Systems." External applications can use those URIs to organize ontologies or tag them.

4.2. Features from AgroPortal inherited from the NCBO BioPortal

The main features offered by the NCBO BioPortal are described in Noy et al. (2009), Whetzel et al. (2011). They include:[16]

Ontology library. The core mission of the AgroPortal is to serve as a one-stop shop for ontology descriptions and files. The portal also allows users to specify the list of ontologies that shall be displayed in their user interface when logged-in. While not replacing source code repository such as for instance GitHub, highly used by the community, the portal stores all ontology versions as they are submitted or automatically pulled, and can display their metadata and differences from one version to the next, although only the latest ontologies are referenced for queries. Ontologies can either be harvested from specified locations, or directly uploaded by users. Ontologies are semantically described (cf. metadata), and a browsing user interface allows to quickly identify, with faceted search, the ontologies of interest based on their descriptions and metadata.

Search across all the ontologies. AgroPortal search service indexes the ontology content (classes, properties and values) with Lucene, and offers an endpoint to search across the ontologies by keyword or identifier. For example, a keyword search on "abiotic factor"[17] will identify the occurrence of this term (or similar terms if none match exactly) in all the ontologies of the portal, and sort the results by relevance to the query and ontology popularity in the portal (number of views) (Noy et al., 2013). For the above search, the first three results are Abiotic factor (CO_715_0000078), Abiotic stress (CO_320:Abiotic_stress), and abiotic stress trait (TO_0000168).

Ontology browsing and content visualization. The ontology 'classes' and 'properties' tab lets users visualize a class or property within is hierarchy, as well as see the related content (labels, definition, mappings, any other relations). An important point is that each

[16] The features of the portal inherited from the NCBO BioPortal are more extensively described in other publications that are referenced here. We provide here only a small summary as well as relevant agronomy related examples. In addition, the documentation of the portal is also available: https://github.com/agroportal/documentation.

[17] http://agroportal.lirmm.fr/search?q = Abiotic%20factor

AgroPortal content page can be accessed by a direct URL, that can be potentially used to dereference an ontology URI. Dereferencing (or resolving) means to obtain a concrete representation of the identified resource (e.g., a web page), for instance, http://agroportal.lirmm.fr/ontologies/EOL/?p = classes&conceptid = http://opendata.inra.fr/EOL/EOL_0000014 directly points to the class 'water salinity' in Environment Ontology for Livestock. For each ontology, a JavaScript widget allowing autocomplete with class names is also automatically generated and can be used by external web applications to facilitate the edition of data fields restricted to ontology concepts.

Ontology versioning. AgroPortal handles versioning through the concept of "submission." Once an "Ontology" (an empty skeleton with minimal metadata) has been added once to the portal, "submission" objects can be attached. A new submission is created every time that ontology is re-submitted by a user, or pulled from its original location URL. Many ontologies are not necessarily maintained in a versioning system which offers a pull URL. It is up to the developer to decide when to manually uploading the new file, thereby creating a new submission (version) in AgroPortal. However, when the ontology is configured with a pull URL, the new ontology will be pulled in automatically (and versioned as a new submission) any night that it has changed. For example, the Matter Transfer Ontology for instance is developed by INRA using the @ Web application (http://pfl.grignon.inra.fr/atWeb).[18] Although only the latest version is indexed and therefore available for searching, browsing and annotation, all the previous versions are downloadable, and a difference comparison can be viewed for each submission.

Ontology mappings. Another key role of AgroPortal is to store mappings (or alignments) between ontologies (Ghazvinian et al., 2009). Indeed, because ontologies' contents overlap, it is crucial to maintain their interconnections—mappings—alongside the ontologies themselves. AgroPortal implements a mapping repository where each class-to-class mapping added to the portal is a first-class citizen and can be: stored, described, retrieved and deleted. The portal automatically creates some mappings when two classes share the same URI or CUI properties,[19] or when they share a common normalized preferred label or synonym. Although basic lexical mapping approaches can be inaccurate and should be used with caution (Faria et al., 2014; Pathak and Chute, 2009), they usually work quite well with the LOOM mapping algorithm used in AgroPortal (Ghazvinian et al., 2009). Other mappings can be explicitly uploaded from external sources, and in that case a mapping is reified as a resource described with provenance information (e.g., automatic or manual, who added it) and one or several tags to classify the mapping (e.g., owl:sameAs, skos:exactMatch, skos:broaderMatch, gold:translation). Such information helps users decide if they want to use these mappings.

Community feedback. While not being a state-of-the-art Web 2.0 social platform for ontologies, the AgroPortal features a few community features (Noy et al., 2009) such as: (i) *Ontology reviews*: for each ontology, a review can be written by a logged-in user from the ontology "Summary" page. It helps keep track of the quality. (ii) *Manual mapping creation*: On each ontology class, a logged-in user can create a mapping to another class (whether the class is inside the AgroPortal, or in the NCBO BioPortal or another resource (cf. next Section)) (Noy et al., 2008). While this is illustrative, and may stimulate propositions, the real strength of the portal comes from using the API to automatically import mappings. (iii) *Notes* can be attached in a forum-like mode to a specific ontology or class, in order to discuss the ontology (its design, use, or evolution) or allow users to propose changes to a certain class (for instance, see http://agroportal.lirmm.fr/ontologies/CO_321/?p = notes). Ontology developers (or any registered users) can subscribe to email notifications to be informed each time user feedback is added to their ontologies of interest.

Ontology-based annotation. AgroPortal features a text annotation service that will identify ontology classes inside any text (Jonquet et al., 2009) and can filter the results per ontologies and UMLS Semantic Types (McCray, 2003).[20] The text annotation service provides a mechanism to employ ontology-based annotation in curation, data integration, and indexing workflow; it has been used to semantically index several data resources such as in the NCBO Resource Index (Jonquet et al., 2011).[21] The workflow is based on a highly efficient syntactic concept recognition tool (using concept names and synonyms) (Dai et al., 2008), and on a set of semantic expansion algorithms that leverage the semantics in ontologies (e.g., is_a relations and mappings). The Annotator is illustrated Fig. 1. It is also used to recommend ontologies for given text input, as described hereafter.

Ontology recommendation. The NCBO (in collaboration with LIRMM & University of Coruña) has recently released a new version of the Recommender system in BioPortal (Martinez-Romero et al., 2017), which has also been installed in AgroPortal. This service suggests relevant ontologies from the parent repository for annotating text data. The new recommendation approach evaluates the relevance of an ontology to biomedical text data according to four different criteria: (1) the extent to which the ontology covers the input data; (2) the acceptance of the ontology in the community; (3) the level of detail of the ontology classes that cover the input data; and (4) the specialization of the ontology to the domain of the input data. This new version of a service originally released in 2010 (Jonquet et al., 2010) combines the strengths of its predecessor with a range of adjustments and new features that improve its reliability and usefulness. To our knowledge, the AgroPortal Recommender is the first ontology recommendation service made for the agronomy community to identify which ontologies are relevant for (i) a given corpus of text or (ii) a list of keywords. For instance, if used with the 'Plant height' text example, from Fig. 1. the service will help users to identify Trait Ontology and multiple sources from the Crop Ontology as relevant for this text.

Register ontology related projects. The AgroPortal provides a project list edited by its users that materialize the ontology-project relation. For instance, the relation between the Planteome project and the six ontologies it uses is described at http://agroportal.lirmm.fr/projects/Planteome, in a format that can be used by AgroPortal to illustrate the ontologies that are most used. This information can then be employed for instance to sort ontologies by number of projects that use them.

In addition, all the previous features are available through two endpoints allowing automatic querying of the content of the portal: (i) a REST web service API (http://data.agroportal.lirmm.fr/

[18] There are 328 submissions as of March 2017: http://data.agroportal.lirmm.fr/ontologies/TRANSMAT/submissions. The latest one is always available under http://data.agroportal.lirmm.fr/ontologies/TRANSMAT/latest_submission

[19] Uniform Resource Identifiers (URIs) are the standard way to identify resources (classes, properties, instances) on the semantic web when using RDF-based languages such as OWL or SKOS. Concept Unique Identifiers (CUIs) are identifiers used in the UMLS Metathesaurus. They are heavily used in the biomedical domain, but not very relevant within AgroPortal, where only two sources (the Semantic Network and the NCBI Taxonomy) are extracted from the UMLS.

[20] This feature originally developed for the NCBO Annotator (Jonquet et al., 2009) allows to filter the annotation results using the upper level 127 UMLS semantic type (http://agroportal.lirmm.fr/ontologies/STY) with which each concept in the UMLS are tagged. Because this was very useful on the NCBO BioPortal, we are considering an equivalent network and mechanism in the AgroPortal.

[21] The 'Resource Index' feature is not used in AgroPortal. Our vision is to accomplish this with the AgroLD partner project.

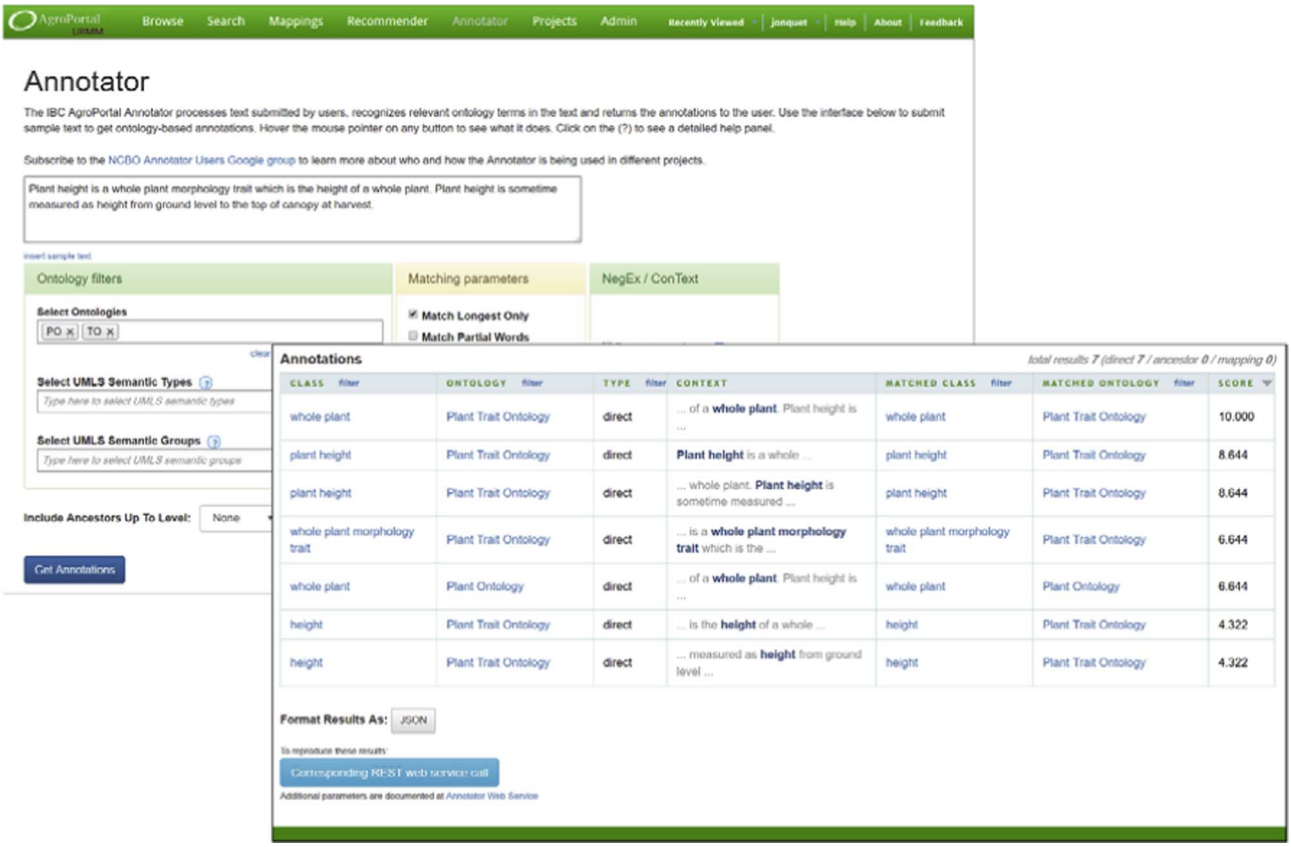


Fig. 1. AgroPortal Annotator with scored results. (web service call: http://services.agroportal.lirmm.fr/annotator?text=Plant height is a whole plant morphology trait which is the height of a whole plant. Plant height is sometime measured as height from ground level to the top of canopy at harvest.&ontologies=PO,TO&longest_only=true &whole_word_only=true&score=cvalue).

documentation) that returns XML or JSON-LD, making it easy to use AgroPortal within any web based application (Whetzel et al., 2011); and (ii) a SPARQL endpoint (http://sparql.agroportal.lirmm.fr/test), which is the standard mechanism to query RDF data (Salvadores et al., 2012).

We also like to point out that by adopting the NCBO technology, including its web service APIs (Whetzel and Team, 2013), an important number of external applications developed by the biomedical semantics community become available at very low cost for the agronomy community because of backward compatibility. This includes spreadsheet annotation tools such as OntoMaton (Maguire et al., 2013) Weboulous (Jupp et al., 2015), RightField (Wolstencroft et al., 2010) and WebSmatch (Coletta et al., 2012; Castanier et al., 2014); Zooma, a tool similar to the Annotator developed by the European Bioinformatics Institute (www.ebi.ac.uk/spot/zooma); the UIMA wrapper to use the Annotator web service in other NLP applications (Roeder et al., 2010); the ontology wrapper OntoCAT (Adamusiak et al., 2010); the Galaxy platform tools (Miñarro-Giménez et al., 2012); the visualization tool FlexViz (Falconer et al., 2009); and finally all the different API clients (Java, Ruby, Perl, etc.) developed by the NCBO (https://github.com/ncbo) or other organizations (e.g. REDCap or Protégé plugins). To some extent, other ontology platforms such as the AberOWL, which features reasoning capabilities that AgroPortal does not yet offer (Slater et al., 2016), can automatically pull content from the AgroPortal.

4.3. New AgroPortal features developped since the beginning of the project

While assuring community support, day-to-day maintenance and monitoring of the portal and keeping it up-to-date with the NCBO technology, we have worked on customizations and specific services. These services target the agronomic community, but that could in some cases be used for any domains. With the vision of collaborative development of BioPortal and AgroPortal, when relevant and possible, we push new features back to the main NCBO code branch where BioPortal users or the appliance itself can benefit. The AgroPortal open source code and documentation are accessible on GitHub: https://github.com/agroportal.

Multilingualism in AgroPortal. In the context of the SIFR project and in consultation with the NCBO, we are working on making BioPortal multilingual (Jonquet et al., 2015). This is still work in progress, although we have already added relevant metadata properties to: (i) identify the natural language in which labels are available; and (ii) link monolingual ontologies to their translations. We have also changed the representation of multilingual translation mappings. For the moment, we have chosen to consider English as the main language of AgroPortal (i.e., the one use to display content as well as indexed for Search, Annotator and Recommender services). Multilingual ontologies (i.e., with labels in multiple languages) are parsed, but only the English content is explicitly used. Non-English monolingual ontologies are attached as “views” of a main ontology that is solely described with metadata (no content). For instance, the French Agroecology Knowledge Management ontology, used in a French collaborative network (http://agroportal.lirmm.fr/ontologies/GECO) is only described with metadata but has attached a specific view (http://agroportal.lirmm.fr/ontologies/GECO-FR) with the real content in French.

Mapping related features. In order to interconnect AgroPortal with the NCBO BioPortal or any other repositories, we have changed the model of AgroPortal mappings to store mappings to ontologies (i) in another instance of the BioPortal technology (‘inter-portal’), (ii) in any ‘external’ resources. Hence, any AgroPortal class can be linked to any class in other knowledge resource (e.g., DBPedia, WordNet, AgroLD) or the NCBO BioPortal itself). Mappings are described with

provenance data and typed with a property from a standard semantic web vocabulary (e.g., OWL, SKOS, GOLD). For instance:

- o The class 'plant organ' in the Plant Ontology has been manually mapped to the 'Plant organ' entity in the DBPedia knowledge base. The mapping tag used is skos:exactMatch which means that the classes represent the same entity, while not supporting a logical substitution (as with owl:sameAs).
- o The class 'biomass' in the Biorefinery ontology has been manually mapped to the class 'Biomass' in MeSH on the NCBO BioPortal, and automatically mapped to the class 'biomass' in the AnaEE Thesaurus.
- o The class 'zooplankton' in the AnaEE Thesaurus has been mapped to 'zooplankton' in the Ontology for MIRNA Target (http://purl.obolibrary.org/obo/OMIT_0015869), which is not available in AgroPortal.

Semantic annotation with scoring. Within the SIFR project we develop new features and natural language based enhancement that target all the Annotator deployments (the NCBO, AgroPortal or SIFR one). For instance, to facilitate the use of annotation for semantic indexing, we have implemented three scoring methods for the Annotator. They are based on term frequency and especially useful with multi-word terms. We demonstrate the results of these new scoring measures in Melzi and Jonquet (2014). For instance, when considering annotating the text:[22] "*Plant height is a whole plant morphology trait which is the height of a whole plant. Plant height is sometime measured as height from ground level to the top of canopy at harvest.*" with the AgroPortal Annotator, the scoring method gives more importance to the concept 'plant height' (score = 8.64) than to the concept 'height' (score = 4.32), whose lexical form is actually more frequent in the text. The user interface of the Annotator is illustrated in Fig. 1.

Ontology formats. We have worked on the full support of different formats such as (i) SKOS (SKOS-XL is not handled yet), which is highly used in agronomy (AnaEE Thesaurus, Agrovoc, CAB Thesaurus and NAL Thesaurus all use SKOS); and (ii) the Crop Ontology Trait Dictionary template v5, adopted for instance by the Breeding API and Crop Ontology (import/export in this format is currently done outside of AgroPortal).

Ontology metadata. To facilitate the ontology identification and selection process, which has been assessed as crucial to enable ontology reuse (Park et al., 2011), we implemented a new metadata model to better support descriptions of ontologies and their relations, respecting recent metadata specifications, vocabularies, and practices used in the semantic web community (Xiang et al., 2011). We reviewed the most common and relevant vocabularies (23 in total) to describe metadata for ontologies, including Dublin Core, VoID, Ontology Metadata Vocabulary, and the Data Catalog Vocabulary. We then grouped those properties into a unified and simplified model of 127 properties (distilled from an initial list of 346 properties that will be parsed by the portal)[23] that includes the 45 properties originally offered by the NCBO BioPortal, and describe all the new properties with standard vocabularies.[24] This gives us, for example, a model to describe the type of the semantic resource uploaded to the portal (for example, thesaurus, ontology, taxonomy, or terminology). Our work provided three important new features for AgroPortal (Toulet et al., 2016):

- o Once an ontology is uploaded, AgroPortal automatically extracts most of the ontology metadata if they are included in the original file, and automatically populates some of them if possible (e.g., metrics, endpoints, links, examples). Ontology developers can manually update those extracted or calculated values if desired. In addition, we have entirely redesigned AgroPortal's ontology submission page to facilitate editing the metadata. Whenever possible, the user interface facilitates the selection of the metadata values, while in the backend those values are stored with standard URIs. For instance, the user interface will offer a pop-up menu to select the relevant license (CC, BSD, etc.) while the corresponding URI will be taken from the RDFLicense dataset (http://rdflicense.appspot.com). Knowledge organization systems types are taken from the KOS Types Vocabulary from the Dublin Core initiative.[25] An example using the OntoBiotope ontology metadata page in AgroPortal is shown in Fig. 2.
- o AgroPortal ontology browse page (Fig. 3) offers three additional ways to filter ontologies in the list (content, natural language, formality level) as well as three new options to sort this list. We believe these new features facilitate the process of selecting relevant ontologies.
  - o We have begun facilitating the comprehension of the agronomical ontology landscape by displaying diagrams and charts about all the ontologies on the portal (average metrics, most used tools, leading contributors & organization, and more). We have created a new AgroPortal 'landscape' page that displays metadata "by property" –as opposed as "by ontology" as in Fig. 2 (http://agroportal.lirmm.fr/landscape).

For each ontology available and uploaded in the portal, we collaborate with the ontology developers to extensively describe their metadata. Information is generally found either in other registries (e.g., LovInra, VEST Registry, the OBO Foundry) or identified in the publication, web site, documentation, etc. found about the ontologies. With these curated metadata, all users can confidently select and review ontologies; any submission of the ontology can include more authoritative and more complete metadata, available to any user including the original provider, and for other linked open data users and applications; and AgoPortal's users can better understand the landscape of ontologies in the agronomy and related domains.

5. Driving agronomic use case results

Now that AgroPortal has been extensively presented, we focus on the results of each use case, and illustrate the value added by this portal and its semantic content.

5.1. Agronomic Linked Data (AgroLD)

The OWL versions of the ontologies available in AgroPortal were retrieved from that single repository. Although AgroPortal is not the main original location for these ontologies (they are accessible on the OBO Foundry and Cropontoloy.org) it was convenient to find them all in one place, and to use a unique and consistent API. Plus, we also used the AgroPortal Annotator web service to annotate more than 50 datasets and produced 22% additional triples, which were validated manually (Fig. 4). Building such an annotation service for all these ontologies was one of the driving needs for AgroPortal. Encoding the original data in RDF allowed us to establish an annotation for every appropriate case, using owl:sameAs relations, between the data element (e.g., Protein in the SouthGreen database) defined with a new URI (http://www.southgreen.fr/agrold/resource/Protein) and an ontology term (e.g., the term 'polypeptide' in the Sequence Ontology (http://purl.obolibrary.org/obo/SO_0000104). Note that we have decided to use owl:sameAs in this case as the resources are logically equivalent and this is a common practice in linked open data to

[22] Two appended definitions from the Trait Ontology and from the Crop Ontology.

[23] https://github.com/agroportal/documentation/tree/master/metadata

[24] For instance, the call http://data.agroportal.lirmm.fr/ontologies/PR/latest_submission?display = all will display the JSON-LD format of all the metadata properties (populated or not) for the Protein Ontology.

[25] http://wiki.dublincore.org/index.php/NKOS_Vocabularies (ANSI/NISO Z39.19-2005).

Fig. 2. AgroPortal's Ontolgy metadata page for ONTOBIOTOPE (http://agroportal.lirmm.fr/ontologies/ONTOBIOTOPE). The red box corresponds to the new metadata fields added in AgroPortal ontology model extracted by the portal, or provided by the adminstrators or by the ontology developers. (For interpretation of the references to colour in this figure legend, the reader is referred to the web version of this article.)

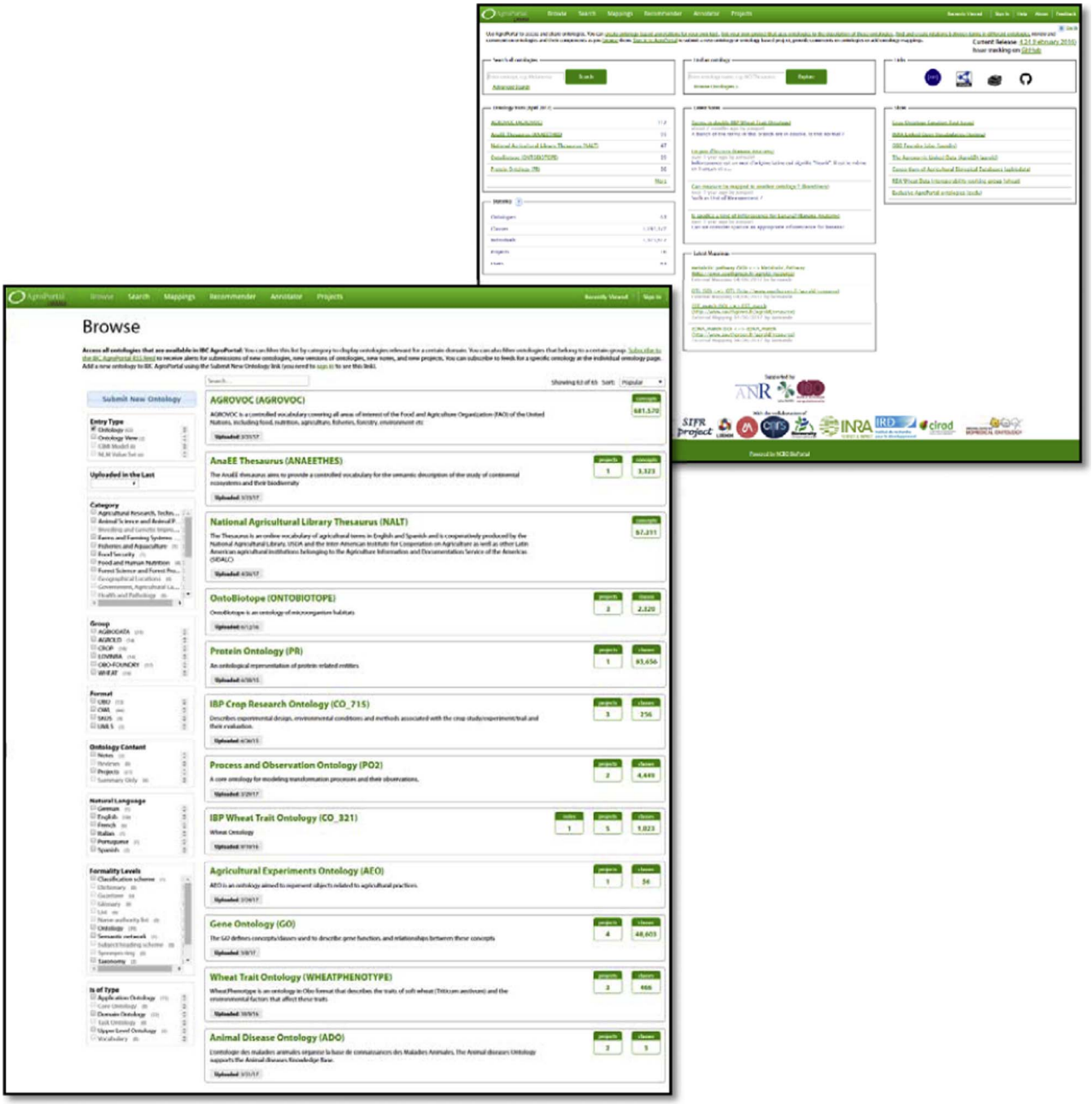


Fig. 3. Screeshots from the AgroPortal user interface (http://agroportal.lirmm.fr). The welcome page (back) provides a rapid overview of the content of the portal and enables a user to quickly search for and in ontologies. The browse ontology page (front) provides the list of ontologies and offer multiple sorting or faceted filtering of this list to facilitate the identification of the ontologies of interest.

interlink datasets; similar annotations have been made for properties using owl:equivalentProperty or rdfs:subPropertyOf (when an equivalent property did not exist). Now that AgroPortal handles 'external mapping' as described in Section 4.3, we have been able to upload all our annotations (to 23 classes and 21 properties) to fully connect the concepts from the different ontologies, and create annotations, directly within AgroPortal.[26]

As a result, AgroLD has incorporated the data from various databases (Table 4), and produced 37 million RDF triples (Venkatesan et al., 2015). The data source selection followed the needs and priorities of the IBC project's work-package 5. It included important data sources such as GOA, Gramene, Oryza Tag Line, and GreenPhylDB. AgroLD can now gather genomic and phenotypic information to answer biological questions such as: "find proteins involved in plant disease resistance and high grain yield traits." Such queries would be hard or impossible to resolve without the appropriate ontologies integrated to support the conclusion. The reader may refer to http://agrold.org/sparqleditor.jsp for more examples of queries in AgroLD.

5.2. RDA Wheat Data Interoperability (WDI) working group

We created and maintain explicit sub-parts within AgroPortal called

[26] The previous example ('polypeptide' in SO) is available here in the mapping tab: http://agroportal.lirmm.fr/ontologies/SO?p = classes&conceptid = http%3A%2F%2Fpurl.obolibrary.org%2Fobo%2FSO_0000104

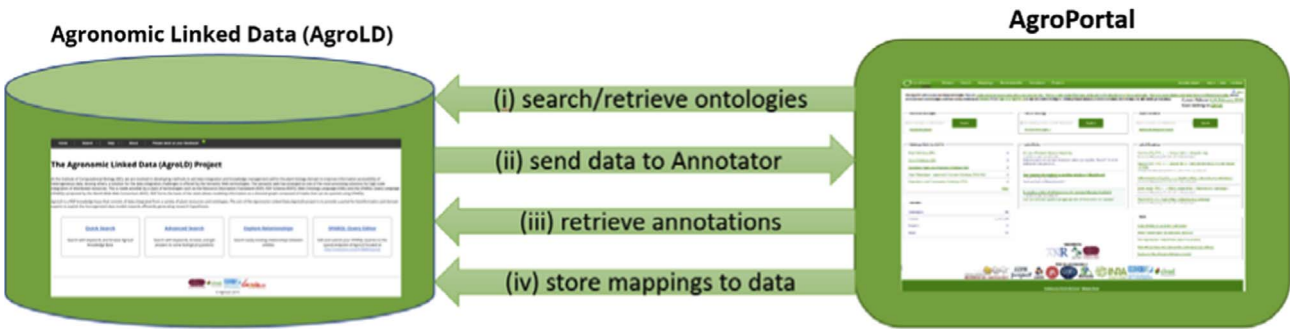


Fig. 4. Interaction between AgroPortal and AgroLD. (i) AgroPortal provides a unique endpoint to retrieve heterogenous ontologies; (ii) AgroLD's annotation pipleline send data to the AgroPortal Annotator and (iii) retrieves annotations with ontology terms used to build AgroLD; finally (iv) AgroPortal offers a link from the ontologies to data stored in AgroLD with the 'inter portal' mapping mechanism.

Table 4
Plant species and data sources in AgroLD. The number of tuples gives an idea of the number of elements we have annoated from the data sources and the number of RDF triples produced. The crops and ontologies are refered as: R = rice, W = wheat, A = Arabidopsis, S = sorghum, M = maize GO = Gene Ontology, PO = Plant Ontology, TO = Plant Trait Ontology, EO = Environment Ontology, SO = Sequence Ontology, CO = Crop Ontoloy (specific trait ontologies).

Data sources	URL s	# tuples	Crops	Ontologies used	# triples produced
GO associations	geneontology.org	1160 K	R, W, A, M, S	GO, PO, TO, EO	2700 K
Gramene	gramene.org	1718 K	R, W, M, A, S	GO, PO, TO, EO	5172 K
UniprotKB	uniprot.org	1400 K	R, W, A, M, S	GO, PO	10000 K
OryGenesDB	orygenesdb.cirad.fr	1100 K	R, S, A,	GO, SO	2300 K
Oryza Tag Line	oryzatagline.cirad.fr	22 K	R	PO, TO, CO	300 K
TropGeneDB	tropgenedb.cirad.fr	2 k	R	PO, TO, CO	20 K
GreenPhylDB	greenphyl.org	100 K	R, A	GO, PO	700 K
SniPlay	sniplay.southgreen.fr	16 K	R	GO	16000 K
TOTAL					37000 K

slices.[27] The wheat slice in AgroPortal (http://wheat.agroportal.lirmm.fr) allows the community to share common definitions for the words they utilize to describe and annotate data, which in turn makes the data more machine-readable and interoperable. Furthermore, each slice enables ontology developers to make their ontologies more visible to targeted agronomic research communities; as of today, AgroPortal's Wheat group contains 20 of the 23 ontologies identified by the WDI.[28] Each ontology has been carefully described (with licenses, authority, availability, and so on), and a new metadata property (omv:endorsedBy) is used to show the ontology's endorsement by the WDI working group.

This work has been reported in the WDI's set of guidelines for wheat data description (http://ist.blogs.inra.fr/wdi) (Dzalé-Yeumo et al., 2017), and used since then as a reference to identify and select ontologies related to wheat. Among AgroPortal's registered users, a dozen are members of the RDA WDI working group. In the future, the slice will be maintained/managed by the WheatIS consortium to organize new wheat-related ontologies and store the alignments between them. AgroPortal's adoption by the WDI working group leveraged several advanced features of the platform as customized by the AgroPortal team. The result directly enhanced the community's processes and capabilities, provided customized access to information of particular interest to this community, and achieved wide uptake in the working group.

5.3. INRA Linked Open Vocabularies (LovInra)

To augment the visibility of INRA's semantic resources, and achieve their mapping to resources within and external to INRA, the institute has chosen AgroPortal to publish and host INRA's resources and encourage adoption of semantic web standards. If a semantic resource is declared on the LovInra service, it is immediately uploaded and fully described on AgroPortal. Resources that are not on the LovInra service can be directly uploaded by their developers to the portal, an important consideration for such a big organization. AgroPortal assigns the new resources to the correct group and slice, and properly tags them (SKOS vocabularies, OWL/SKOS termino-ontological resources, or OBO/OWL ontologies).

The LovInra group/slice contains 16 ontologies relating to process modeling, biotopes, animal breeding, and plant phenotypes. AgroPortal has become a major element of the LovInra service and is heavily encouraged and supported by INRA. It has started to play a key resource role allowing the group's users to: (i) have a comprehensive view of the portal's ontologies (topics, types, community, etc.); (ii) quickly find a resource, and understand its content and structure by browsing it and annotating documents; (iii) discover additional vocabularies that could be used; and (iv) have access to projects linked to vocabularies, and understand how they were created or used by the projects, possibly exchanging shared experience or insights.

5.4. The Crop Ontology project

Currently, the AgroPortal hosts 19 crop-specific trait ontologies developed within the Crop Ontology project: Wheat, Rice, Cassava, Groundnut, Chickpea, Banana, Sweet potato, Cowpea, Soybean, Lentil, Pigeon pea, Sorghum, Pear millet, Maize, Groundnut, Castor bean, Mungbean, and Cassava. Additional ontologies will be integrated in the future with the help of the crop ontology curators. Similarly to the

[27] Slices are a mechanism supported by the platform to allow users to interact (both via API or UI) only with a subset of ontologies in AgroPortal. If browsing the slice, all the portal features will be restricted to the chosen subset, enabling users to focus on their specific use cases. On AgroPortal, slices and groups are synchronized, so every group (described Section 4.1) has a corresponding slice displaying only the ontologies from that group.

[28] Among the missing ones are, CAB Thesaurus, that we are currently working on integrating; CheBI that we have decided not to upload yet; and Wheat Inra Phenotype Ontology (that is currently being merged with CO_321).

LovInra or WDI use cases, these ontologies are grouped within the portal and can be browsed in a dedicated slice (http://crop.agroportal.lirmm.fr). Parsers for specific trait template have been developed, and in the future any of this community's formats (OBO, OWL, and CSV) shall be used to import and export trait ontologies directly within AgroPortal.[29]

Moreover, in the context of the Planteome project (www.planteome.org), the alignment (or mapping) of terms within and across different plant related ontologies have been created: both within the crop ontologies themselves (in different crop branch) or with other reference ontologies commonly used in plant biology (e.g., PO, TO, EO). In the future, AgroPortal will formally store the alignments between all these ontologies.[30]

Finally, hosting ontologies on AgroPortal offers new functionalities to the crop ontology community such as versioning, an open SPARQL endpoint, community notes, and the annotation service, while still supporting the uses of the current web site.[31] For instance, new traits or mappings between them can be suggested directly by breeders using AgroPortal's community features, while not directly impacting the original ontology. Each time a suggestion is made to an ontology, the breeders interested in the corresponding crop can be notified of the suggestions and comments of their peers.

5.5. GODAN Map of Agri-Food Data Standards

The GODAN Action project wanted to build a broadly scoped global map of standards while leveraging detailed information and content about them that could be maintained in an ontology or vocabulary. To achieve this, the new map of standards was built on top of the existing VEST Registry, but added bidirectional mechanisms linking the VEST Registry with AgroPortal. The combined system automatically imports resource descriptions from the AgroPortal into the VEST, and links records from the VEST back to the AgroPortal entries, in order to provide access to the AgroPortal content and related services. The new registry, called *Map of Agri-Food Data Standards* (http://vest.agrisemantics.org), was released in 2016 under two umbrellas: the GODAN Action project, and the new RDA AgriSemantics working group,[32] which launched at the end of 2016. The Map of Standards leverages the AgroPortal's new metadata model and application programming interface to populate the entries in the Map using a single web service call. In addition to searching by metadata, the AgroPortal's Recommender will help the agronomy community identify ontologies or vocabularies of interest.

The synchronization and interlinking of the two platforms is for the moment semi-automatic, with the content of AgroPortal being regularly imported into the global map. Users can register or edit the description of a vocabulary in the Map, and if the vocabulary is in a compatible format, they are offered, the option to add the vocabulary directly into AgroPortal. In the future, this process will be fully automatized.

6. Discussion

6.1. General reflection on research scenarios supported by AgroPortal

AgroPortal (like the NCBO BioPortal before it) adopted a vision where multiple knowledge artifacts are made available in a common place (though not combined), and cast to a common model. While doing so, the portal arguably limits the full power of ontologies, constraining their use to features supported by the common model. We see two general scenarios of use for our portal:

> The portal provides basic ontology library services for users with a "vertical need" —those who want to do very precise things (e.g., reasoning, using specific relations) using only suitable ontologies (developed by the same communities and in the same format). Such users may just use the portal to find and download ontologies, and work in their own environment.
> The portal provides many semantic services (for examples, lexical analysis, search, text annotation, and use of hierarchical knowledge) to users with "horizontal needs" —those who wants to work with a wide range of ontologies and vocabularies useful in their domain but developed by different communities, overlapping and in different formats. Such users greatly appreciate the unique endpoints (web application and programmatic for REST and SPARQL queries) offered by the portal under a simplified common model.

We believe there are existing resource to address the first need in agronomy (e.g., OBO Foundry, FAIRSharing, VEST registry), although without containing all the relevant ontologies and vocabularies. However, we argue the second need is unmet by any of the available platforms. If we want semantic resources like ontologies and vocabularies to achieve widespread adoption, we must facilitate their use for non-ontological experts who still want to use multiple heterogeneous semantic resources.

6.2. Implementation of the requirements

As presented and illustrated on examples, most of the requirements listed in Section 3 have been addressed at least partially thanks to the original BioPortal features (e.g., requirements #1-#6, #8, #10, #15, #16, #18), our new implementations (#5, #7, #11, #15), and our applying the platform to the community needs (#1, #10, #11, #17, #18). Some requirements are not yet completely achieved and/or evaluated, for instance:

> (#4) The AgroPortal Annotator has been used by the AgroLD use case, but not by other ones. We have not yet evaluated the capability of the service to automatically identify entities such as plant phenotypes in text.
> (#8) Automatically generating mappings is an important issue for a portal on ontologies. Although it is convenient to have some simple lexical mappings automatically generated by AgroPortal with the LOOM algorithm (Ghazvinian et al., 2009), we find that this is not enough to correctly interlink the multiple vocabularies and ontologies developed by the community. We are integrating other state-of-the-art ontology matchers such as YAM++ (Ngo and Bellahsene, 2012) as well as designing specific mapping curation interfaces. At the same time, identifying and harvesting into AgroPortal the mappings already produced by the community is a huge task, not yet begun.
> (#9) We have not automatically linked databases of annotated agronomical data using ontology concepts (from within AgroPortal). While the original BioPortal has the NCBO Resource Index (Jonquet et al., 2011), we plan to rely on external annotated resources such as AgroLD (Venkatesan et al., 2015) to interlink with data. To store this information, we will build on our rich mapping model in AgroPortal as presented Section 4.3. As another example, being part of the map of standards will allow ontologies in AgroPortal to link directly to

[29] Most of these conversions are still achieved outside of AgroPortal. The automatically generated CSV output format is not yet compliant with the Crop Ontology trait template (v5).

[30] For instance, something to capture that plant height for wheat (CO_321:0000024) is somehow linked to the general plant height trait (TO_0000207) that is itself a morphology trait (TO:0000398). This work is ongoing, and the data is not yet publicly released.

[31] In the future, to offer to breeders a simple and customized interface while avoiding duplication effort, we will consider serving the Crop Ontology website use cases by directly accessing AgroPortal's backend through the REST API.

[32] https://www.rd-alliance.org/groups/agrisemantics-wg.html

datasets that use them such as the CIARD RING directory (http://ring.ciard.net) (Pesce et al., 2011), as that was previously indexed with some of the VEST content. The CIARD RING can be queried via SPARQL or REST API and the links between vocabularies and datasets can therefore be retrieved by any system. Such a feature, has been requested and will be among the next features of AgroPortal. In the long-term vision, AgroPortal will directly query the CIARD RING, AgroLD, or any relevant data sources like Bio2RDF or Planteome, so that a user browsing ontologies can get direct access to the data to which these ontologies link.

(#12) Although community feedback is an important aspect for working group and communities, we have not successfully engaged yet our user groups to add reviews, notes, or comments about the ontologies. A complete rethinking of this issue is a future challenge for AgroPortal.

(#13) The roadmap to make the technology fully multilingual has been identified, but not yet fully implemented.

(#15) AgroPortal can be used as a destination for dereferenced URIs. In the future, we shall discuss these strategic questions with our collaborators.

6.3. Future and perspectives

Considering the need for a repository of ontologies for agronomy, food, plant sciences, and biodiversity, we expect broad community adoption of the AgroPortal. The endorsement of associated partners (IRD, CIRAD, INRA, IRSTEA) illustrates the impact and interest not just in France, but also internationally (e.g., FAO, Bioversity International, IC-FOODS consortium, NCBO, Planteome, RDA working groups). More recently, two other RDA working groups (Rice Data Interoperability[33] and AgriSemantics[34]) have expressed interest in using AgroPortal as a backbone for data integration and standardization.

In the future, we will identify more potential users for the portal and support new research scenarios. For instance, within the RDA AgriSemantics WG, we are interested in using AgroPortal to host the future Global Agricultural Concept Scheme (GACS) (Baker et al., 2016), which will result from the integration and alignment of Agrovoc, NAL Thesaurus and CAB Thesaurus. The portal is considered by the GACS working group as a candidate to host the three source vocabularies (it already includes two of them), as well as the GACS itself. GACS beta version 3.1 is currently available in AgroPortal, but no specific customization has been performed. In addition, we will be offering our services to these projects:

the new IC-FOODS project (International Center for Food Ontology, Operability, Data & Semantics - www.ic-foods.org) that will be developing ontologies related to food, nutrition, eating behaviors (Musker et al., 2016);

ecologists developing the Thesaurus of Plant characteristics (Garnier et al., 2017);

the French IRESTA organization, to facilitate the use of ontologies in the design of the future government-led open data repository for agriculture project (AgGate).[35]

To foster interest in agronomy and the semantic web and identify potential AgroPortal applications, we launched in 2016 a series of AgroHackathons (www.agrohackathon.org) that focused among other things on AgroPortal and AgroLD. Finally, in the next future, we plan to achieve a community survey evaluation to capture the feedback of our community, review the requirements, and drive the future directions of the project.

7. Conclusion

In this paper we have presented AgroPortal, an open vocabulary and ontology repository for agronomy. We have discussed five use cases already using the portal to support their work on data interoperability, and demonstrated that beyond these use cases the portal offers services of value to the broader community. The thematic boundaries of the portal are evolving (agriculture also includes animals, and is strongly related to environmental science), and over time the community will communicate what they expect to find in such a repository.

The community outreach challenge of such a project is huge. It involves identifying already existing resources, whether already shared or not, encouraging their developers to make them available, and finally harvesting them into the single ontology repository, capable of providing many services across the heterogeneous content. We recognize that this challenge was highly facilitated by previous important efforts such as the NCBO BioPortal, OBO Foundry, Planteome, and Crop Ontology projects. In addition, we are conscious that by adopting an open library approach, knowledge "conflicts" or redundancies as well as convergences and consolidations will appear. We believe the AgroPortal will help the scientific community to fully understand these issues, and address them as appropriate.

The technological challenges of such a project are also huge; therefore, we have built upon technology previously developed in the biomedical domain. We see here an opportunity to capitalize technology and scientific outcomes of the last twelve years in a closely related domain. We illustrated in the context of five important driving agronomic use cases how AgroPortal can enable new science for the community developing and using agronomical ontologies and vocabularies worldwide. In addition, the AgroPortal platform offers a terrain for pursuing important informatics and semantic web issues, such as semantic annotation, multilingual ontologies, metadata description, ontology engineering and alignment, and ontology recommendation, and will.

Ultimately, we believe AgroPortal provides powerful services, standards, and information that will greatly facilitate the adoption of open data in agriculture and benefit the extended agronomic community, the semantic web and data science communities, and the biomedical community that in many ways laid the groundwork that AgroPortal now leverages.

Acknowledgment

This work is partly achieved within the Semantic Indexing of French biomedical Resources (SIFR – www.lirmm.fr/sifr) project that received funding from the French National Research Agency (grant ANR-12-JS02-01001), the European Union's Horizon 2020 research and innovation programme under the Marie Sklodowska-Curie grant agreement No 701771, the NUMEV Labex (grant ANR-10-LABX-20), the Computational Biology Institute of Montpellier (grant ANR-11-BINF-0002), as well as by the University of Montpellier and the CNRS. We also thank the National Center for Biomedical Ontologies for their help and time spent with us in deploying the AgroPortal.

Author contributions

CJ conceived of the project, provided the scientific direction and led the writing of this manuscript. VE & AT respectively implemented/maintained the portal and managed the content with help of the community. JG and MAM helped and gave directions in realizing the project in collaboration with NCBO, and JG provided extensive final review and editing. Then, EA, SA, MAL, EDY, VP & PL respectively presented each of the use cases. All authors declare no conflict of interest and approved the final manuscript.

[33] https://rd-alliance.org/groups/rice-data-interoperability-wg.html
[34] https://rd-alliance.org/groups/agrisemantics-wg.html
[35] https://www.economie.gouv.fr/files/files/PDF/rapport-portail-de-donnees-agricoles.pdf

Appendix

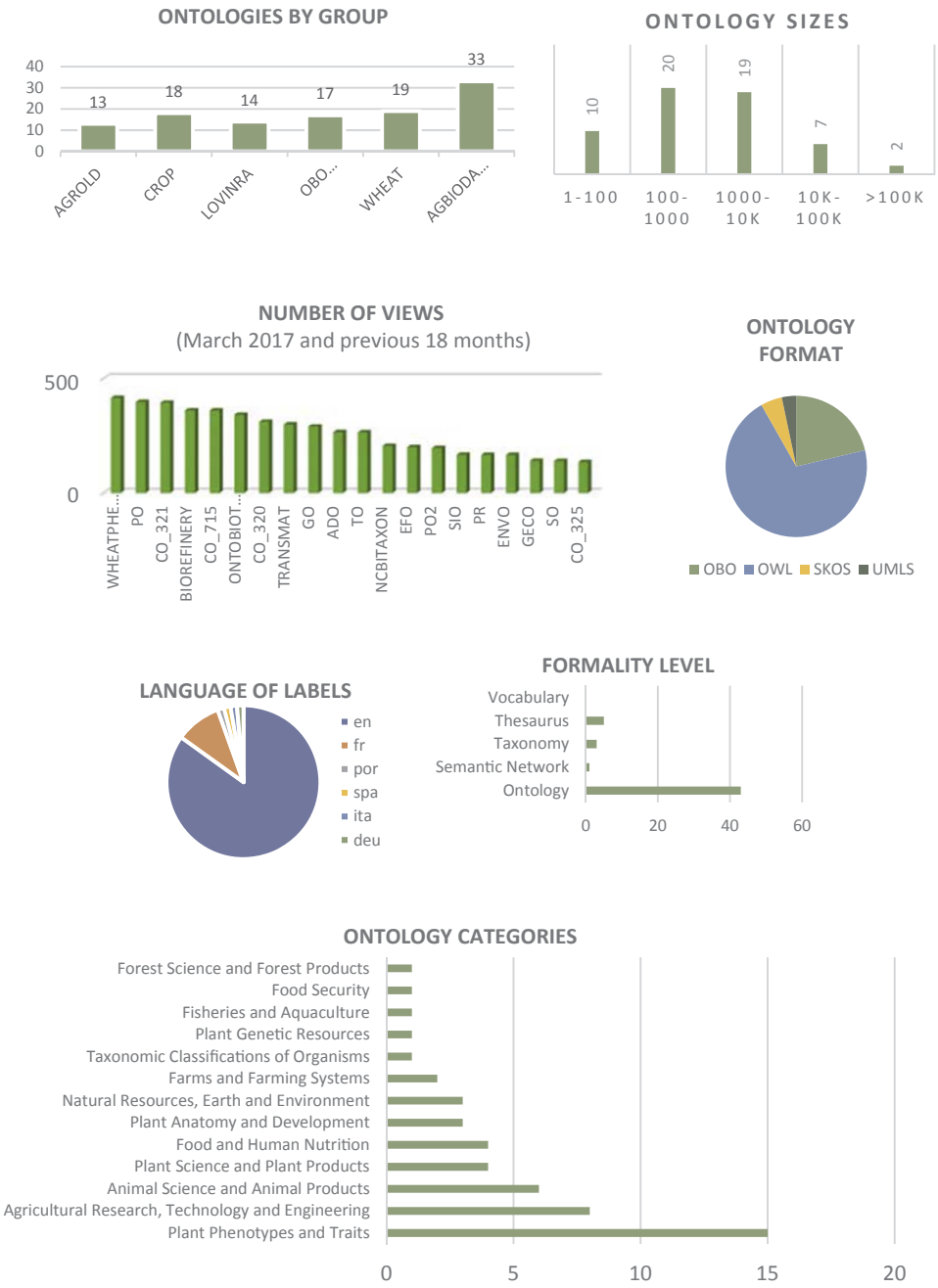


Fig. 5. AgroPortal public ontology analytics (May 2017). Updated versions of these statistics are automatically generated from AgroPortal's new metadata model, and made available on its Landscape page (http://agroportal.lirmm.fr/landscape).

References

Goble, C., Stevens, R., 2008. State of the nation in data integration for bioinformatics. Biomed. Inf. 41, 687–693.

Rubin, D.L., Shah, N.H., Noy, N.F., 2008. Biomedical ontologies: a functional perspective. Brief. Bioinform. 9 (1), 75–90.

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., et al., 2000. Gene ontology: tool for the unification of biology. Nat. Genet. 25 (1), 25–29.

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., Shrestha, R., Yamazaki, Y., Jaiswal, P., 2012. The plant ontology as a tool for comparative plant anatomy and genomic analyses. Plant Cell Physiol. 54, e1.

Shrestha, R., Arnaud, E., Mauleon, R., Senger, M., Davenport, G.F., Hancock, D., Morrison, N., Bruskiewich, R., McLaren, G., 2010. Multifunctional crop trait ontology for breeders' data: field book, annotation, data discovery and semantic enrichment of the literature. AoB Plants, vol. 2010, May.

Buttigieg, P.L., Morrison, N., Smith, B., Mungall, C.J., Lewis, S.E., 2013. The environment ontology: contextualising biological and biomedical entities. Biomed. Semantics 4, 43.

Devare, M., Aubert, C., Laporte, M.-A., Valette, L., Arnaud, E., Buttigieg, P.L., 2016. Data-driven agricultural research for development - a need for data harmonization via semantics. In: Jaiswal, P., Hoehndorf, R. (Eds.), 7th International Conference on Biomedical Ontologies, ICBO'16, vol. 1747 of CEUR Workshop Proceedings, Corvallis, Oregon, USA, pp. 2, August.

Garnier, E., Stahl, U., Laporte, M.-A., Kattge, J., Mougenot, I., Kühn, I., Laporte, B., Amiaud, B., Ahrestani, F.S., Bönisch, G., Bunker, D.E., Cornelissen, J.H.C., Díaz, S., Enquist, B.J., Gachet, S., Jaureguiberry, P., Kleyer, M., Lavorel, S., Maicher, L., Pérez-Harguindeguy, N., Poorter, H., Schildhauer, M., Shipley, B., Violle, C., Weiher, E., Wirth, C., Wright, I.J., Klotz, S., 2017. Towards a thesaurus of plant characteristics: an ecological contribution. Ecology 105, 298–309.

Griffiths, E., Brinkman, F., Buttigieg, P.L., Dooley, D., Hsiao, W., Hoehndorf, R., 2016. FoodON: a global farm-to-fork food ontology - the development of a universal food vocabulary. In: Jaiswal, P., Hoehndorf, R., (Eds.), 7th International Conference on

Biomedical Ontologies, ICBO'16, vol. 1747 of CEUR Workshop Proceedings, Corvallis, Oregon, USA, pp. 2, August.
Musker, R., Lange, M., Hollander, A., Huber, P., Springer, N., Riggle, C., Quinn, J.F., Tomich, T.P., 2016. Towards designing an ontology encompassing the environment-agriculture-food-diet-health knowledge spectrum for food system sustainability and resilience. In: Jaiswal, P., Hoehndorf, R. (Eds.), 7th International Conference on Biomedical Ontologies, ICBO'16, vol. 1747 of CEUR Workshop Proceedings, Corvallis, Oregon, USA, pp. 5, August.
Hughes, L.M., Bao, J., Hu, Z.-L., Honavar, V., Reecy, J.M., 2014. Animal trait ontology: The importance and usefulness of a unified trait vocabulary for animal species. Anim. Sci. 86, 1485–1491.
Meng, X.-X., 2012. Special issue – agriculture ontology. Integrative Agriculture, vol. 11, pp. 1, May.
Walls, R.L., Deck, J., Guralnick, R., Baskauf, S., Beaman, R., Blum, S., Bowers, S., Buttigieg, P.L., Davies, N., Endresen, D., Gandolfo, M.A., Hanner, R., Janning, A., Krishtalka, L., Matsunaga, A., Midford, P., Morrison, N., Tuama, Éamonn Ó., Schildhauer, M., Smith, B., Stucky, B.J., Thomer, A., Wieczorek, J., Whitacre, J., Wooley, J., 2014. Semantics in support of biodiversity knowledge discovery: an introduction to the biological collections ontology and related ontologies. PLoS One 9, 13.
Wang, Y., Wang, Y., Wang, J., Yuan, Y., Zhang, Z., 2015. An ontology-based approach to integration of hilly citrus production knowledge. Comput. Electron. Agric. 113, 24–43.
Lousteau-Cazalet, C., Barakat, A., Belaud, J.-P., Buche, P., Busset, G., Charnomordic, B., Dervaux, S., Destercke, S., Dibie, J., Sablayrolles, C., Vialle, C., 2016. A decision support system for eco-efficient biorefinery process comparison using a semantic approach. Comput. Electron. Agric. 127, 351–367.
Lehmanna, R.J., Reichea, R., Schiefera, G., 2012. Future internet and the agri-food sector: State-of-the-art in literature and research. Comput. Electron. Agric. 89, 158–174.
Jaiswal, P., 2011. Plant Reverse Genetics: Methods and Protocols, ch. Gramene Database: A Hub for Comparative Plant Genomics. Humana Press, pp. 247–275.
Sachit Rajbhandari, J.K., 2012. The AGROVOC concept scheme – a walkthrough. Integrative, Agriculture 11, 694–699.
d'Aquin, M., Noy, N.F., 2012. Where to publish and find ontologies? a survey of ontology libraries. Web Semantics 11, 96–111.
Bodenreider, O., 2004. The unified medical language system (UMLS): integrating biomedical terminology. Nucleic Acids Res. 32, 267–270.
Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L.J., Eilbeck, K., Ireland, A., Mungall, C.J., Consortium, T.O., Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.-A., Scheuermann, R.H., Shah, N.H., Whetzel, P.L., Lewis, S., 2007. The OBO foundry: coordinated evolution of ontologies to support biomedical data integrationNat. Biotechnol. 25, 1251–1255.
Larmande, P., Arnaud, E., Mougenot, I., Jonquet, C., Libourel, T., Ruiz, M., (Eds.), 2013. Proceedings of the 1st International Workshop on Semantics for Biodiversity, Montpellier, France, May.
Baker, T., Caracciolo, C., Jaques, Y., (Eds.), 2015. Report on the Workshop "Improving Semantics in Agriculture, (Rome, Italy), Food and Agriculture Organization of the UN, July.
Noy, N.F., Shah, N.H., Whetzel, P.L., Dai, B., Dorf, M., Griffith, N.B., Jonquet, C., Rubin, D.L., Storey, M.-A., Chute, C.G., Musen, M.A., 2009. BioPortal: ontologies and integrated data resources at the click of a mouse. Nucleic Acids Res. 37, 170–173.
Whetzel, P.L., Noy, N.F., Shah, N.H., Alexander, P.R., Nyulas, C., Tudorache, T., Musen, M.A., 2011. BioPortal: enhanced functionality via new web services from the national center for biomedical ontology to access and use ontologies in software applications. Nucleic Acids Res. 39, 541–545.
Whetzel, P.L., Team, N., 2013. NCBO technology: powering semantically aware applications. Biomed. Semantics 49 4S1.
Ding, Y., Fensel, D., 2001. Ontology library systems: the key to successful ontology re-use. In: 1st Semantic Web Working Symposium, SWWS'01, Stanford, CA, USA, pp. 93–112, CEUR-WS.org, August.
Baclawski, K., Schneider, T., 2009. The open ontology repository initiative: Requirements and research challenges. In: Tudorache, T., Correndo, G., Noy, N., Alani, H., Greaves, M., (Eds.), Workshop on Collaborative Construction, Management and Linking of Structured Knowledge, CK'09, vol. 514 of CEUR Workshop Proceedings, Washington, DC., USA, pp. 10, CEUR-WS.org, October.
Sansone, S.-A., Rocca-Serra, P., Field, D., Maguire, E., Taylor, C., Hofmann, O., Fang, H., Neumann, S., Tong, W., Amaral-Zettler, L., Begley, K., Booth, T., Bougueleret, L., Burns, G., Chapman, B., Clark, T., Coleman, L.-A., Copeland, J., Das, S., de Daruvar, A., de Matos, P., Dix, I., Edmunds, S., Evelo, C.T., Forster, M.J., Gaudet, P., Gilbert, J., Goble, C., Griffin, J.L., Jacob, D., Kleinjans, J., Harland, L., Haug, K., Hermjakob, H., Sui, S.J.H., Laederach, A., Liang, S., Marshall, S., McGrath, A., Merrill, E., Reilly, D., Roux, M., Shamu, C.E., Shang, C.A., Steinbeck, C., Trefethen, A., Williams-Jones, B., Wolstencroft, K., Xenarios, I., Hide, W., 2012. Toward interoperable bioscience data. Nat. Genet. 44, 121–126.
Pesce, V., Geser, G., Protonotarios, V., Caracciolo, C., Keizer, J., 2013. Towards linked agricultural metadata: directions of the agINFRA project. In: 7th Metadata and Semantics Research Conference, AgroSem track, Thessaloniki, Greece, pp. 12, November.
Xiang, Z., Mungall, C., Ruttenberg, A., He, Y., 2011. Ontobee: a linked data server and browser for ontology terms. In: Bodenreider, O., Martone, M.E., Ruttenberg, A., (Eds.), 2nd International Conference on Biomedical Ontology, ICBO'11, vol. 833 of CEUR Workshop Proceedings, Buffalo, NY, USA, p. 3, July.
Côté, R.G., Jones, P., Apweiler, R., Hermjakob, H., 2006. The ontology lookup service, a lightweight cross-platform tool for controlled vocabulary queries. BMC Bioinf. 7, 7.
Hoehndorf, R., Slater, L., Schofield, P.N., Gkoutos, G.V., 2015. Aber-OWL: a framework for ontology-based data access in biology. BMC Bioinf. 16, 1–9.
Vandenbussche, P.-Y., Atemezing, G.A., Poveda-Villalón, M., Vatant, B., 2014. Linked open vocabularies (LOV): a gateway to reusable semantic vocabularies on the web. Semantic Web.
Till, M., Kutz, O., Codescu, M., 2014. Ontohub: A semantic repository for heterogeneous ontologies, In: Theory Day in Computer Science, DACS'14, (Bucharest, Romania), p. 2, September.
Graybeal, J., Isenor, A.W., Rueda, C., 2012. Semantic mediation of vocabularies for ocean observing systems. Comput. Geosci. 40, 120–131.
Ong, E., Xiang, Z., Zhao, B., Liu, Y., Lin, Y., Zheng, J., Mungall, C., Courtot, M., Ruttenberg, A., He, Y., 2016. Ontobee: a linked ontology data server to support ontology term dereferencing, linkage, query and integration. Nucleic Acids Res. 45, D347–D352.
Jonquet, C., Shah, N.H., Musen, M.A., 2009. The open biomedical annotator. In: American Medical Informatics Association Symposium on Translational BioInformatics, AMIA-TBI'09, San Francisco, CA, USA, pp. 56–60, March.
Jonquet, C., LePendu, P., Falconer, S., Coulet, A., Noy, N.F., Musen, M.A., Shah, N.H., 2011. NCBO resource index: ontology-based search and mining of biomedical resources, web semantics. In: 1st Prize of Semantic Web Challenge at the 9th International Semantic Web Conference, ISWC'10, Shanghai, China, vol. 9, pp. 316–324, September.
Salvadores, M., Alexander, P.R., Musen, M.A., Noy, N.F., 2013. BioPortal as a dataset of linked biomedical ontologies and terminologies in RDF. Semantic Web 4 (3), 277–284.
Rueda, C., Bermudez, L., Fredericks, J., 2009. The MMI ontology registry and repository: a portal for marine metadata interoperability. In: MTS/IEEE Biloxi - Marine Technology for Our Future: Global and Local Challenges, OCEANS'09, Biloxi, MS, USA, pp. 6, October.
D.A., Pouchard, L., Huhns, M., 2012. Lessons learned in deploying a cloud-based knowledge platform for the ESIP Federation. In: American Geo-physical Union Fall Meeting, poster session, San Francisco, USA, December.
Jonquet, C., Annane, A., Bouarech, K., Emonet, V., Melzi, S., 2016. SIFR BioPortal: Un portail ouvert et générique d'ontologies et de terminologies biomédicales françaises au service de l'annotation sémantique. In: 16th Journées Francophones d'Informatique Médicale, JFIM'16, Genève, Suisse, pp. 16, July.
Jonquet, C., Emonet, V., Musen, M.A., 2015. Roadmap for a multilingual BioPortal. In: In: Gracia, J., McCrae, J., Vulcu, G. (Eds.), 4th Workshop on the Multilingual Semantic Web, MSW4'15, vol. 1532. CEUR Workshop Proceedings, Portoroz, Slovenia, pp. 15–26.
Matteis, L., Chibon, P., Espinosa, H., Skofic, M., Finkers, R., Bruskiewich, R., Arnaud, E., 2015. Crop ontology: vocabulary for crop-related concepts. In: In: Larmande, P., Arnaud, E., Mougenot, I., Jonquet, C., Libourel, T., Ruiz, M. (Eds.), 1st International Workshop on Semantics for Biodiversity, vol. 1. CEUR Workshop Proceedings, Montpellier, France, pp. 37–46.
Noy, N.F., Tudorache, T., Nyulas, C., Musen, M.A., 2010. The ontology life cycle: Integrated tools for editing, publishing, peer review, and evolution of ontologies. In: AMIA Annual Symposium, Washington DC, USA, pp. 552–556, November.
Jaiswal, P., Cooper, L., Elser, J.L., Meier, A., Laporte, M.-A., Mungall, C., Smith, B., Johnson, E.K., Seymour, M., Preece, J., Xu, X., Kitchen, R.S., Qu, B., Zhang, E., Arnaud, E., Carbon, S., Todorovic, S., Stevenson, D.W., 2016. Planteome: A resource for Common Reference Ontologies and Applications for Plant Biology. In: 24th Plant and Animal Genome Conference, PAG'16, San Diego, USA, January.
Carbon, S., Ireland, A., Mungall, C.J., Shu, S., Marshall, B., Lewis, S., 2009. AmiGO: online access to ontology and annotation data. Bioinformatics 25 (2), 288–289.
Schmachtenberg, M., Bizer, C., Paulheim, H., 2014. Adoption of the linked data best practices in different topical domains. In: Mika, P., Tudorache, T., Bernstein, A., Welty, C., Knoblock, C., Vrandecic, D., Groth, P., Noy, N., Janowicz, K., Goble, C., (Eds.), 13th International Semantic Web Conference, ISWC'14, vol. 8796 of Lecture Notes in Computer Science, Riva del Garda, Italy, Springer, pp. 245–260, October.
Belleau, F., Nolin, M.-A., Tourigny, N., Rigault, P., Morissette, J., 2008. Bio2RDF: towards a mashup to build bioinformatics knowledge systems. Biomed. Inf. 41, 706–716.
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., Jenkinson, A.M., 2014. The EBI RDF platform: linked open data for the life sciences. Bioinformatics 30, 1338–1339.
Groth, P., Loizou, A., Gray, A.J., Goble, C., Harland, L., Pettifer, S., 2014. API-centric linked data integration: the open PHACTS discovery platform case study. Web Semantics 29, 12–18.
Venkatesan, A., Hassouni, N.E., Philippe, F., Pommier, C., Quesneville, H., Ruiz, M., Larmande, P., 2015. Exposing French agronomic resources as linked open data. In: 8th Semantic Web Applications and Tools for Life Sciences International Conference, SWAT4LS'15, vol. 546 of CEUR Workshop Proceedings, Cambridge, UK, pp. 205–207, December.
Hassani-Pak, K., Zorc, M., Taubert, J., Rawlings, C., 2013. QTLNetMiner - candidate gene discovery in plant and animal knowledge networks. In: 21st Plant & Animal Genome Conference, poster session, San Diego, USA, pp. P0980, January.
Dzalé-Yeumo, E., et al., 2017. Developing data interoperability using standards: A wheat community use case, F1000 Research, 6–1843, October 2017. (In preparation).
Shrestha, R., Matteis, L., Skofic, M., Portugal, A., McLaren, G., Hyman, G., Arnaud2, E., 2012. Bridging the phenotypic and genetic data useful for integrated breeding through a data annotation using the Crop Ontology developed by the crop communities of practice. Frontiers Physiol. 3.
Coletta, R., Castanier, E., Valduriez, P., Frisch, C., Ngo, D., Bellahsene, Z., 2012. Public data integration with Websmatch. In: Raschia, G., Theobald, M., (Eds.), 1st International Workshop on Open Data, WOD'12, Nantes, France, pp. 5–12, ACM, May.
Castanier, E., Jonquet, C., Melzi, S., Larmande, P., Ruiz, M., Valduriez, P., 2014. Semantic

Sempéré et al. GigaScience (2016) 5:25
DOI 10.1186/s13742-016-0131-8

GigaScience

TECHNICAL NOTE Open Access

Gigwa—Genotype investigator for genome-wide analyses

Guilhem Sempéré[1,2*], Florian Philippe[3], Alexis Dereeper[2,4], Manuel Ruiz[2,5,6,7], Gautier Sarah[2,8] and Pierre Larmande[2,3,6,9]

Abstract

Background: Exploring the structure of genomes and analyzing their evolution is essential to understanding the ecological adaptation of organisms. However, with the large amounts of data being produced by next-generation sequencing, computational challenges arise in terms of storage, search, sharing, analysis and visualization. This is particularly true with regards to studies of genomic variation, which are currently lacking scalable and user-friendly data exploration solutions.

Description: Here we present Gigwa, a web-based tool that provides an easy and intuitive way to explore large amounts of genotyping data by filtering it not only on the basis of variant features, including functional annotations, but also on genotype patterns. The data storage relies on MongoDB, which offers good scalability properties. Gigwa can handle multiple databases and may be deployed in either single- or multi-user mode. In addition, it provides a wide range of popular export formats.

Conclusions: The Gigwa application is suitable for managing large amounts of genomic variation data. Its user-friendly web interface makes such processing widely accessible. It can either be simply deployed on a workstation or be used to provide a shared data portal for a given community of researchers.

Keywords: Genomic variations, VCF, HapMap, NoSQL, MongoDB, SNP, INDEL, Web interface

Findings

Background

With the advent of next-generation sequencing (NGS) technology, thousands of new genomes of both plant and animal organisms have recently become available. Whole exome and genome sequencing, genotyping-by-sequencing and restriction site-associated DNA sequencing (RADseq) are all becoming standard methods to detect single-nucleotide polymorphisms (SNPs) and insertions/deletions (indels), in order to identify causal mutations or study the associations between genetic variations and functional traits [1–4]. As a result, huge amounts of gene sequence variation data are accumulating in numerous species-oriented projects, such as 3000 rice genomes [5] or 1001 *Arabidopsis* genomes [6, 7]. In this context, the Variant Call Format (VCF) [8] has become a convenient and standard file format for storing variants identified by NGS approaches.

VCF files may contain information on tens of millions of variants, for thousands of individuals. Having to manage such significant volumes of data involves considerations of efficiency with regard to the following aspects:

1. Filtering features. Such data can be processed by applications like VCFtools [8], GATK [9], PyVCF [10], VariantAnnotation [11] or WhopGenome [12] to query, filter and extract expertized datasets for day-to-day research. However, these tools are limited to command line or programmatic application programming interfaces (APIs) targeted at experienced users, and are not suitable for non-bioinformaticians.
2. Storage performance. Working with flat files is not an optimal solution in cases where scientists need to establish comparisons across projects and/or take metadata into account. The use of relational

* Correspondence: guilhem.sempere@cirad.fr
[1]UMR InterTryp (CIRAD), Campus International de Baillarguet, 34398, Montpellier, Cedex 5, France
[2]South Green Bioinformatics Platform, 1000 Avenue Agropolis, 34934 Montpellier, Cedex 5, France
Full list of author information is available at the end of the article

databases is still widely prevalent within the range of more integrated approaches. However, such solutions have limitations when managing big data [13]. In computational environments with large amounts of heterogeneous data, the NoSQL database technology [14, 15] has emerged as an alternative to traditional relational database management systems. NoSQL refers to non-relational database management systems designed for large-scale data storage and massively parallel data processing. During the past 5 years, a number of bioinformatics projects have been developed based on NoSQL databases such as HBase [16, 17], Hadoop [18–20], Persevere [21], Cassandra [22] and CouchDB [23].

3. Sharing capabilities. This aspect is clearly best addressed by providing client/server-based applications, which enable multiple users to work on the same dataset without the need to replicate it for each user. There is, as yet, a considerable lack of web applications able to handle the potentially huge genotyping datasets that are emerging from mass genotyping projects, and which would enable biologists to easily access, query and analyze data online.
4. Graphical visualization. A number of solutions have been developed for the graphical visualization of genomic variation datasets. Some of these have been integrated into data portals associated with specific projects (e.g. OryzaGenome [24], SNP-Seek [25]) and are, therefore, only relevant to a particular community. Generic tools also exist (e.g. vcf.iobio [26, 27], JBrowse [28]) and may be built upon to create more versatile applications.

The Gigwa application, the name of which stands for 'Genotype investigator for genome-wide analyses', aims to take account of all of these aspects. It is a web-based, platform-independent solution that feeds a MongoDB [29] NoSQL database with VCF or HapMap files containing billions of genotypes, and provides a web interface to filter data in real time. In terms of visualization, the first version includes only an online density chart generator. However, Gigwa supplies the means to export filtered data in several popular formats, thus facilitating connectivity with many existing visualization engines.

Application description

A single instance of the Gigwa application is able to display data from multiple databases, which can be chosen from a drop-down menu at the top of the page. A database may syndicate any source of genotyping data as long as the variant positions are provided on the same reference assembly. Gigwa supports work on a single project at a time (although a project may be divided into several runs, in which case new data connected to existing individuals are seen as additional samples). Project selection may be changed from within the action panel (Fig. 1) that sits to the right-hand side of the screen. This panel also enables the launch of searches, toggles the availability of browsing and exporting functions (because limiting the initial approach to the counting of results saves time), configures and launches the export, checks the progress of ongoing operations, and can terminate them if required.

The variant-filtering interface (top of Fig. 2) is both compact and intuitive. In the top-left corner of this panel, three lists allow multiple-item selection of variation types (e.g. SNPs, indels, structural variants), individuals and reference sequences. More specific filters can be incorporated to refine searches using combinations of the following parameters:

- 'Genotypes' - this filter makes it possible to retrieve only variant positions that respect a specified genotyping pattern when considering selected individuals. If no individuals are selected, the application takes them all into account. A dozen predefined options (e.g. all same, at least one heterozygous) are available, covering those cases that are most frequently meaningful.
- 'Minimum per-sample genotype quality' and 'Minimum per-sample read depth' - these individual-based filters may be used, in the case of data from VCF files, to set thresholds on the quality (GQ) and depth (DP) fields assigned to genotypes [30]. Individuals that do not meet these criteria are subsequently treated as missing data.
- 'Authorized missing data ratio' - this filter allows a maximum threshold of acceptable missing data among selected individuals to be defined. Its default value is 100 %, that is, accepting all data.
- 'Minor allele frequency' (MAF) - this filter retains only the variant positions for which the MAF calculated on selected individuals falls in the

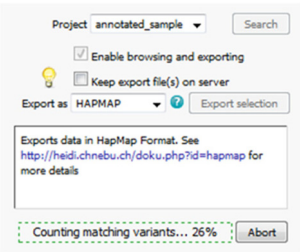


Fig. 1 Action panel enabling project selection, progress indication, abort and export functionalities

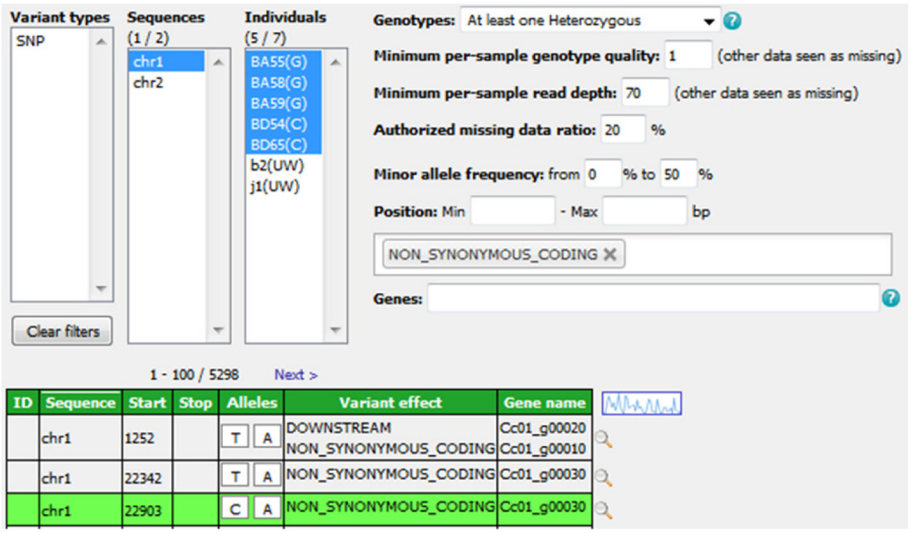


Fig. 2 Variant-filtering interface, allowing search criteria definition and result browsing

specified range (by default, 0–50 %). It is only applicable to bi-allelic markers.

- 'Number of alleles' - this filter allows specification of the number of known alleles the targeted variants are expected to have.
- 'Position' - this filter restricts the search to variants located in a given range of positions in relation to the reference.
- SnpEff widgets - these allow additional filtering on variant effects and gene names for data originating from VCF files that have been annotated with SnpEff [31]. The application automatically detects such additional data and is able to handle both types of annotation field, that is, 'EFF' (SnpEff versions prior to 4.1) and 'ANN' (SnpEff versions from 4.1 onwards).

Matching variants are displayed in paginated form (see bottom of Fig. 2) after application of the filters. Results are listed in a sortable table that provides the main attributes, namely ID (when provided in the input file), reference sequence, start and stop positions, alleles, variant effect and gene name (the latter two only being displayed if available). In addition, the user can focus on a specific position and display variant details, including selected individuals' genotypes, using the magnifier at the end of each row. These details appear in a dialogue (Fig. 3) that, for each run in the selected project, provides:

- additional variant-level attributes or annotations (global attributes related to the variant), on the left of the screen;

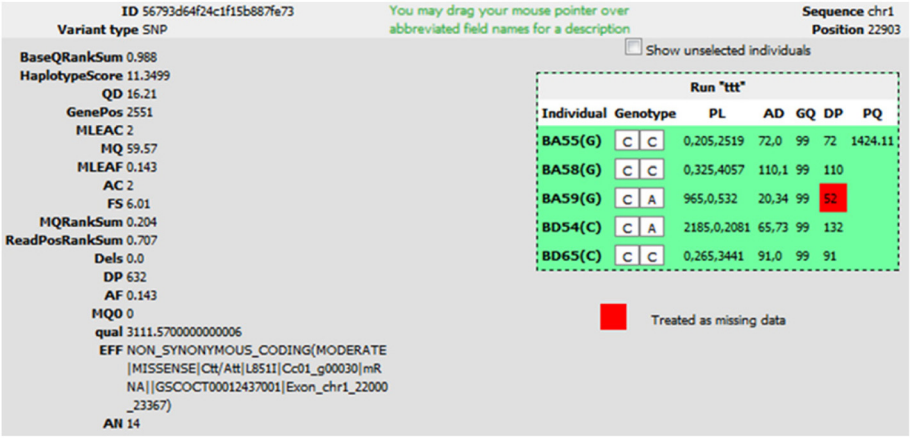


Fig 3 Variant detail dialogue, providing variant metadata and genotype-level information

- on the right of the screen, a box indicating each individual's genotype, along with genotype-level attributes (e.g. depth, quality). A checkbox allows the display of genotypes for unselected individuals. Any GQ and DP values that are below specified thresholds, and have thus led to a genotype being considered as missing, are highlighted with a red background.

Data export and visualization

The Gigwa application offers seven standardized formats (VCF, Eigenstrat, GFF3, BED, HapMap, DARwin and PLINK) in which to export filtered results in compressed files. Export is individual-based. Thus, if the data selection includes several samples that belong to the same individual, only one genotype per variant is exported. If these genotypes are inconsistent, the one most frequently found is selected. If there is no most frequently found genotype, one is picked at random.

Where data that originated from VCF files is being re-exported in the same format, the application takes phased genotypes into account: a procedure was implemented to maintain phasing information (i.e. haplotype estimation) in the database and recalculate it at export time, even if intermediate positions had been filtered out.

Exports can be directed either to the client computer or to a temporary URL on a web server, thus making the dataset instantly shareable, for example, with Galaxy [32]. Such links remain active for a week. In addition, when applied to the VCF format, this 'export to URL' feature provides the means for users to view selected variants in their genomic context in a running instance of the Integrative Genomics Viewer (IGV) [33].

The current selection may also be directed to an online, interactive, density chart viewer. The variant distribution of each sequence may then be observed, with the ability to filter on variation type, and these figures can also be exported in various file formats (i.e. PNG, JPEG, PDF and SVG).

Technical insights

Third-party software involved

MongoDB [29] was chosen as the storage layer for several reasons: its complex query support, its scalability, its open-source nature and its proactive support community. The server application was developed in Java and takes advantage of several Spring Framework modules [34] (e.g. Spring Data). The client interface was designed using Java Server Pages (JSP) and jQuery [35]. Some import and export procedures make use of the SAMtools HTSJDK API v1.143 [36]. The density visualization tool was implemented using the HighCharts Javascript library [37].

Data structure

The data model for storing genotyping information, defined using Spring Data documents, is shared with the WIDDE application [21] and allows a single database to hold genotypes from multiple runs of multiple projects. This model is marker-oriented and mainly relies on two basic document types: *VariantData*, which embeds variant-level information (e.g. position, marker type); *VariantRunData*, which contains genotyping data along with possible metadata.

A collection named *taggedVariants* is not tied to a model object because its documents only contain variant IDs. However, it serves an important purpose by providing dividers ('landmarks') that partition the entire collection of variants into evenly sized chunks. These chunks are then used when querying directly on the *VariantRunData* collection (i.e. without a preliminary filter on variant features) to split the query into several subqueries, which confers several advantages (see Querying strategy below).

Less significant model objects include *GenotypingProject*, which keeps track of elements used to rapidly build the interface (e.g. distinct lists of sequence names and variant types involved in the project), and *DBVCFHeader*, which simply stores the contents of headers for runs imported in VCF format.

Querying strategy

When the *Search* button is clicked, the values selected in the search-interface widgets are passed to the server application. They may then be used to count and/or browse matching variants.

The first time a given combination of filters is invoked, the *count* procedure is launched to establish the number of variants that match the combination. This result is then cached in a dedicated collection so that whenever a user subsequently repeats the same search, the result will be available instantly.

Once the *count* result has been displayed to the user, if the 'Enable browsing and exporting' box is checked, a second request is sent to the server, invoking the *find* procedure that eventually provides paginated, detailed variant information in the form of a comprehensive table.

In general, serving such requests (*count* or *find*) may be divided into two consecutive steps:

- a simple, preliminary query of variant features (variant type, sequence, start position), which is applied to indexed fields and therefore executes quickly;
- the main aggregation query, which is split into several partial queries aimed at running in simultaneous threads on evenly sized variant chunks of the *VariantRunData* collection. This technique not only improves performance, but also allows

Gigwa to provide a progress indicator and the facility to terminate a run before it has finished. The method used for dividing the main query depends on whether or not a preliminary filter was executed beforehand. If it was, the application holds a subset of variant IDs as a consequence, which it uses to split the data using MongoDB's *$in* operator in each sub-query. Otherwise, the contents of the *taggedVariants* collection are used, in conjunction with the *$lte* (less than or equal) and *$gt* (greater than) operators, to define the limits of each sub-query's chunk.

Summary of features

Gigwa's value resides in the following features:

- Support for large genotyping files with up to several million variants
- Responsive queries even in the case of a local deployment
- Intuitive graphical user interface allowing the definition of precise queries in a few clicks
- Filtering on functional annotations
- Ability to abort running queries
- Display of query progress
- Support of multiple data sources for a single instance
- A multi-user mode which enables both public and private access to databases to be defined
- Support for incremental data loading
- Support for seven different export formats
- Easy connection with IGV for integration within a consistent genomic context
- No loss of phasing information when provided (VCF format only)
- Support for haploid, diploid and polyploid data
- Online variant density viewing.

The Gigwa application therefore represents a very efficient, versatile and user-friendly tool for users with standard levels of expertise in web navigation to explore large amounts of genotyping data, identify variants of interest and export subsets of data in a convenient format for further analysis. We believe that its large panel of undoubtedly useful features will make Gigwa an essential tool in the increasingly complex field of genomics.

Benchmarking

In order to assess Gigwa's performance, we conducted benchmarks against comparable applications.

Hardware used

All tests were run on an IBM dx360 M2 server with:

- two quad-core CPUs (Core-i7 L5520 at 2.26 GHz);
- 36 GB RAM (DDR3 at 1333 MHz);
- 250-GB SATA2 hard drive.

Dataset selection

As our base dataset, we chose to use the CoreSNP dataset from the 3000 Rice Genomes Project [5, 38], which at the time of download (v2.1) contained genotypes for 3000 individuals on 365,710 SNPs. This dataset was first converted to a 4.09-GB VCF file using VCFtools, from which three progressively smaller datasets were then generated by successively dividing the number of variants by ten, i.e. resulting in datasets of 36,571, 3658 and 366 SNPs, respectively.

Benchmark comparisons

We considered it appropriate to compare Gigwa's performance with that of:

1. VCFtools (v0.1.13) [12];
2. A MySQL (v5.6.28) [39] implementation of a standard relational database model with indexes on appropriate fields. Corresponding queries were implemented as stored procedures, and both these and the database schema are provided as supplementary material within the supporting data.

In addition, the opportunity was taken to evaluate the relative performance of the currently available storage solutions offered by MongoDB v3.0.6, i.e. the newly introduced WiredTiger (WT) storage engine, configured with three different compression levels (none, *snappy* and *zlib*), and the original MMapv1 storage engine.

Therefore, each of the benchmarking plots generated contains six series: VCFtools, MySQL, Gigwa-MMapv1, Gigwa-WT-none, Gigwa-WT-snappy and Gigwa-WT-zlib. MongoDB queries were launched via the Gigwa interface because, internally, the application splits them into a number of partial, concurrent queries.

Benchmark queries

Two kinds of queries that we considered representative were executed as benchmarks on each dataset for each tool:

- Location-based query: a query counting variants located in a defined region of a chromosome (chromosome 3, 1 Mbp to 5 Mbp).
- Genotype-based query: a query counting variants exhibiting a given MAF range (10 to 30 %) on the first 2000 individuals (out of 3000).

All benchmarks were executed three times, except for the MAF queries in Gigwa where the different

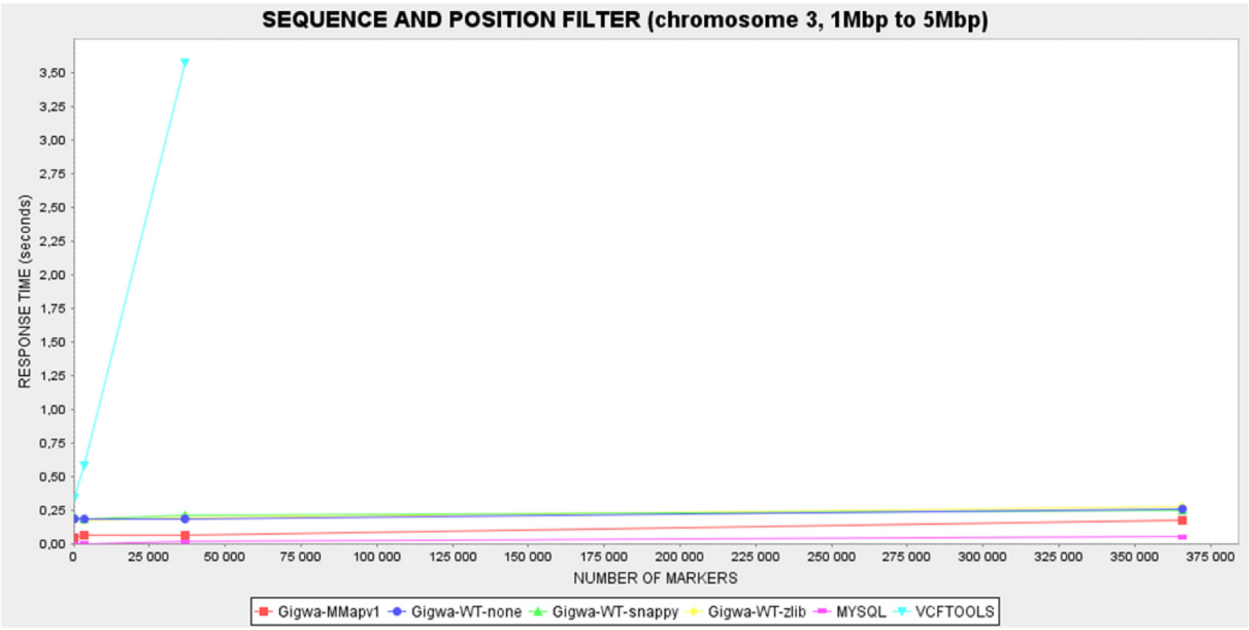


Fig. 4 Response-time plot by tool for first benchmarked query (location-based filter). VCFtools is by far the slowest option

MongoDB configurations gave response times showing high degrees of heterogeneity. In order to establish more distinction between them, these benchmarks were therefore executed 12 times. In all cases, average response times were calculated and then reported through graphical plots. The caching system implemented in Gigwa was disabled for the duration of the benchmarking.

Benchmark results

In the case of the location-based filter benchmark (Fig. 4), the MySQL solution was the fastest, with response times that were negligible on the smallest dataset, and never more than 0.05 s on the largest. In comparison, Gigwa queries were less responsive but still remained fairly fast, never taking more than 0.3 s on the largest dataset. However, VCFtools proved so much

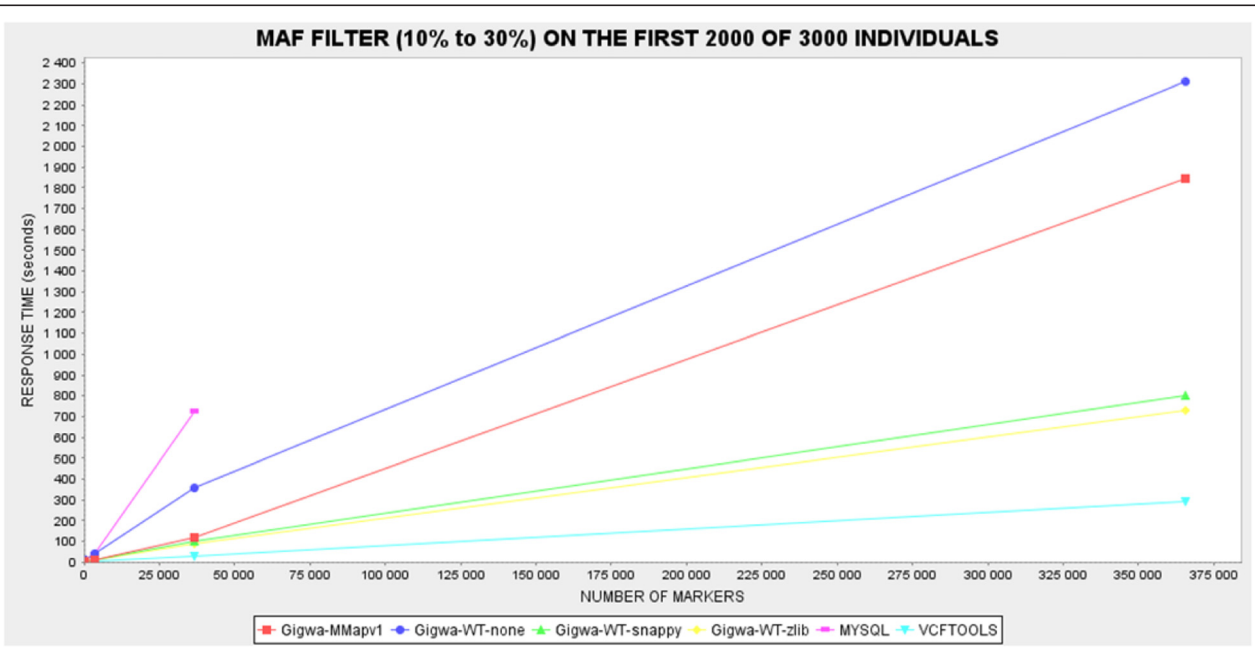


Fig. 5 Response-time plot by tool for second benchmarked query (MAF filter). MySQL is by far the slowest option

slower than all the other alternatives benchmarked that we had to exclude its last record for the plot to remain readable. This difference can be explained by the fact that database engines typically take advantage of pre-built indexes that lead directly to results, whereas VCFtools has to scan the entire file. Relational databases are usually the most efficient for this kind of simple query because their indexing mechanisms have been optimized over decades.

In the case of the MAF filter benchmark (Fig. 5), the fastest solution was VCFtools, followed by the two most compressed MongoDB databases (Gigwa-WT-zlib and Gigwa-WT-snappy), and then by the two least compressed MongoDB databases (Gigwa-MMapv1 and Gigwa-WT-none). The MySQL engine performed so poorly here that it was considered unnecessary to run the longest query on it. In practice, the type of analysis involved in this particular benchmark requires that all stored positions be scanned. VCFtools excels here because it is a C++ program working on flat files, which means that the time needed to access each record is negligible, whereas database engines need to obtain/deflate objects before manipulating them. In contrast to the situation seen in the first filter benchmark, a significant difference in performance emerged here between the various storage solutions offered by MongoDB. There is more room in this benchmark for performance distinctions because memory consumption becomes more crucial when executing a multi-step aggregation pipeline rather than a simple index count. WiredTiger applies compression to indexes, which leaves more memory available for other tasks, thus increasing performance. In addition, WiredTiger is known to perform better than MMapv1 on multi-threaded queries, which are being used by Gigwa.

Thus, Gigwa configured with WiredTiger-snappy (or WiredTiger-zlib in the case of constraints on disk space) appears to be an excellent compromise, being the only solution that responds in a reasonable time to both kinds of query. Furthermore, although it was beyond the scope of this benchmark, we should mention that the greatly reduced storage space required by both WiredTiger-snappy and WiredTiger-zlib, when compared with that required by MMapv1, provides an additional justification for choosing WiredTiger in most cases.

Conclusions

We developed Gigwa to manage large genomic variation data derived from NGS analyses or high-throughput genotyping. The application aims to provide a user-friendly web interface that makes real-time filtering of such data, based on variant features and individuals' genotypes, widely accessible. Gigwa can be deployed either in single-user mode or in multi-user mode, with credentials and permissions allowing fine-grained control of access to connected databases.

We ran benchmarks on two kinds of queries - variant-oriented and genotype-oriented - to compare Gigwa's performance with that of both VCFtools and a standard MySQL model. Each of these latter tools performed best in one benchmark but by far the worst in the other. Gigwa, when configured with the WiredTiger storage engine and either the *snappy* or *zlib* compression level, appeared as an excellent compromise, performing almost as well as the best solution in both benchmarks.

Future versions of Gigwa will include a RESTful API to allow external applications to interact with Gigwa and query data in a standardized manner, as well as additional visualization tools and a Docker [40] package aimed at distributing the tool as a solution capable of functioning in platform-as-a-service (PaaS) [41] mode. Further benchmarks will be conducted to evaluate the application's performance in a distributed environment using MongoDB's sharding functionality.

Availability and requirements

- **Project name:** Gigwa
- **Project home page:** http://www.southgreen.fr/content/gigwa
- **Operating system(s):** Platform-independent
- **Programming language:** Java & MongoDB
- **Requirements:** Java 7 or higher, Tomcat 7 or higher, MongoDB 3 or higher
- **License:** GNU GPLv3
- **Restrictions to use for non-academics:** None

Additional file

Additional file 1: Gigwa, Genotype investigator for genome-wide analyses. Provides the MySQL scripts used for benchmarking and guidelines on how to import data into Gigwa and configure its access for existing users. (DOCX 95 kb)

Acknowledgements

This project was funded by UMR DIADE and Agropolis Fondation under the reference ID ARCAD 0900-001. The authors thank the South Green Platform team for technical support, Sébastien Ravel for testing the annotation filters, Francois Sabot and Mathieu Rouard for meaningful advice, and Aravind Venktesan and Jean-Marc Mienville for careful reading that helped improve the manuscript.

Availability of supporting data

Gigwa's source code is available in South Green's public GitHub repository [42]. Supplementary data, benchmarking material and installation archives can be found in the *GigaScience* GigaDB repository [43].

Authors' contributions

MR provided the original idea. GSempéré designed the application logic. GSempéré and FP implemented the software. AD and GSarah generated the benchmark datasets. PL and GSempéré conducted the benchmarks. AD and

MR performed beta-testing. PL, GSempéré and AD drafted the manuscript. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]UMR InterTryp (CIRAD), Campus International de Baillarguet, 34398, Montpellier, Cedex 5, France. [2]South Green Bioinformatics Platform, 1000 Avenue Agropolis, 34934 Montpellier, Cedex 5, France. [3]UMR DIADE (IRD), 911 Avenue Agropolis, 34934 Montpellier, Cedex 5, France. [4]UMR IPME (IRD), 911 Avenue Agropolis, 34394 Montpellier, Cedex 5, France. [5]UMR AGAP, CIRAD, 34398 Montpellier, Cedex 5, France. [6]Institut de Biologie Computationnelle, Université de Montpellier, 860 Rue de St Priest, 34095 Montpellier, Cedex 5, France. [7]Agrobiodiversity Research Area, International Center for Tropical Agriculture (CIAT), 6713 Cali, Colombia. [8]INRA, UMR AGAP, 34398 Montpellier, Cedex 5, France. [9]INRIA Zenith Team, LIRMM, 161 Rue Ada, 34095 Montpellier, Cedex 5, France.

Received: 12 February 2016 Accepted: 16 May 2016
Published online: 06 June 2016

References

1. Gheyas A, Boschiero C, Eory L, Ralph H, Kuo R, Woolliams J, et al. Functional classification of 15 million SNPs detected from diverse chicken populations. DNA Res. 2015;22(3):205–17.
2. Li X, Buitenhuis AJ, Lund MS, Li C, Sun D, Zhang Q, et al. Joint genome-wide association study for milk fatty acid traits in Chinese and Danish Holstein populations. J Dairy Sci. 2015;98(11):8152–63. Available from: http://www.ncbi.nlm.nih.gov/pubmed/26364108.
3. Shinada H, Yamamoto T, Sato H, Yamamoto E, Hori K, Yonemaru J, et al. Quantitative trait loci for rice blast resistance detected in a local rice breeding population by genome-wide association mapping. Breed Sci. 2015;65(5):388–95. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4671699&tool=pmcentrez&rendertype=abstract.
4. Marcotuli I, Houston K, Waugh R, Fincher GB, Burton RA, Blanco A, et al. Genome wide association mapping for arabinoxylan content in a collection of tetraploid wheats. PLoS One. 2015;10(7):e0132787. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4503733&tool=pmcentrez&rendertype=abstract.
5. The 3000 rice genomes project. The 3,000 rice genomes project. Gigascience. 2014; 3:7. http://dx.doi.org/10.1186/2047-217X-3-7
6. Ossowski S, Schneeberger K, Clark RM, Lanz C, Warthmann N, Weigel D. Sequencing of natural strains of Arabidopsis thaliana with short reads. Genome Res. 2008;18:2024–33. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2593571&tool=pmcentrez&rendertype=abstract.
7. Cao J, Schneeberger K, Ossowski S, Günther T, Bender S, Fitz J, et al. Whole-genome sequencing of multiple Arabidopsis thaliana populations. Nat Genet. 2011;43(10):956–63. Available from: http://www.ncbi.nlm.nih.gov/pubmed/21874002.
8. Danecek P, Auton A, Abecasis G, Albers CA, Banks E, DePristo MA, et al. The variant call format and VCFtools. Bioinformatics. 2011;27(15):2156–8. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3137218&tool=pmcentrez&rendertype=abstract.
9. McKenna A, Hanna M, Banks E, Sivachenko A, Cibulskis K, Kernytsky A, et al. The genome analysis toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 2010;20(9):1297–303. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2928508&tool=pmcentrez&rendertype=abstract.
10. Casbon J. PyVCF - A Variant Call Format Parser for Python. 2012. Available from: https://pyvcf.readthedocs.org/en/latest/INTRO.html
11. Obenchain V, Lawrence M, Carey V, Gogarten S, Shannon P, Morgan M. VariantAnnotation: a bioconductor package for exploration and annotation of genetic variants. Bioinformatics. 2014;30(14):2076–8.
12. Wittelsburger U, Pfeifer B, Lercher MJ. WhopGenome: high-speed access to whole-genome variation and sequence data in R. Bioinformatics. 2015;31(3):413–5. Available from: http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btu636.
13. Bach M, Werner A. In: Nawrat MAM, editor. Innovative control systems for tracked vehicle platforms, vol. 2. Cham: Springer International Publishing; 2014. p. 163–74. Available from: http://link.springer.com/10.1007/978-3-319-04624-2.
14. Gajendran, S.K. A survey on NoDQL databases. University of Illinois; 2012. Available from: http://www.masters.dgtu.donetsk.ua/2013/fknt/babich/library/article10.pdf.
15. Moniruzzaman ABM, Hossain SA. Nosql database: New era of databases for big data analytics-classification, characteristics and comparison. CoRR [Internet]. 2013;6(4):1–14. Available from: http://arxiv.org/abs/1307.0191.
16. O'Connor BD, Merriman B, Nelson SF. SeqWare query engine: storing and searching sequence data in the cloud. BMC Bioinf. 2010;11(12):S2. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3040528&tool=pmcentrez&rendertype=abstract.
17. Wang S, Pandis I, Wu C, He S, Johnson D, Emam I, et al. High dimensional biological data retrieval optimization with NoSQL technology. BMC Genomics. 2014;15(8):S3. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=4248814&tool=pmcentrez&rendertype=abstract.
18. Langmead B, Schatz MC, Lin J, Pop M, Salzberg SL. Searching for SNPs with cloud computing. Genome Biol. 2009;10(11):R134. http://genomebiology.com/2009/10/11/R134.
19. Afgan E, Chapman B, Taylor J. CloudMan as a platform for tool, data, and analysis distribution. BMC Bioinf. 2012;13(1):315. http://www.biomedcentral.com/1471-2105/13/315.
20. Schatz MC. CloudBurst: highly sensitive read mapping with MapReduce. Bioinformatics. 2009;25(11):1363–9. Available from: http://bioinformatics.oxfordjournals.org/cgi/doi/10.1093/bioinformatics/btp236.
21. Russ TA, Ramakrishnan C, Hovy EH, Bota M, Burns GAPC. Knowledge engineering tools for reasoning with scientific observations and interpretations: a neural connectivity use case. BMC Bioinf. 2011;12(1):351. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3176268&tool=pmcentrez&rendertype=abstract.
22. Ye Z, Li S. Arequest skewaware heterogeneous distributed storage systembased on Cassandra. the International Conference on Computer and Management (CAMAN'11). 2011. p. 1–5.
23. Manyam G, Payton M A, Roth J A, Abruzzo L V, Coombes KR. Relax with CouchDB - Into the non-relational DBMS era of bioinformatics. Genomics. Elsevier Inc.; 2012. Available from: http://www.ncbi.nlm.nih.gov/pubmed/22609849. Accessed 19 Dec 2015.
24. Ohyanagi H, Ebata T, Huang X, Gong H, Fujita M, Mochizuki T, et al. OryzaGenome : Genome Diversity Database of Wild Oryza Species Special Online Collection – Database Paper. 2016;0(November 2015):1–7
25. Alexandrov N, Tai S, Wang W, Mansueto L, Palis K, Fuentes RR, et al. SNP-Seek database of SNPs derived from 3000 rice genomes. Nucleic Acids Res. 2015;63(2):2–6.
26. Miller C, Qiao Y, DiSera T, D'Astous B, Marth G. Bam. Iobio: a Web-based, real-time, sequence alignment file inspector. Nat Methods. 2014;11(12):1189.
27. Di Sera TL. vcf.iobio—A visually driven variant data inspector and real-time analysis web application. NEXT GEN SEEK. 2015. Available from: http://vcf.iobio.io/. Accessed 19 Dec 2015.
28. Skinner ME, Uzilov AV, Stein LD, Mungall CJ, Holmes IH. JBrowse: a next-generation genome browser. Genome Res. 2009;19:1630–8. Available from: http://genome.cshlp.org/content/19/9/1630.short.
29. MongoDB Inc. MongoDB. 2015. Available from: https://www.mongodb.org/
30. VCF 4.2 specification. 2015. Available from: https://samtools.github.io/hts-specs/VCFv4.2.pdf
31. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of drosophila melanogaster strain w 1118; iso-2; iso-3. Fly (Austin). 2012;6(June):80–92.
32. Giardine B, Riemer C, Hardison RC, Burhans R, Elnitski L, Shah P, et al. Galaxy: a platform for interactive large-scale genome analysis. Genome Res. 2005;15(10):1451–5. Available from: http://www.ncbi.nlm.nih.gov/pubmed/16169926.
33. Thorvaldsdóttir H, Robinson JT, Mesirov JP. Integrative genomics viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform. 2013;14(2):178–92. Available from: http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=3603213&tool=pmcentrez&rendertype=abstract.
34. Pivotal Software Inc. Java Spring Framework. 2015. Available from: http://projects.spring.io/spring-framework/
35. The jQuery Foundation. JQuery. 2015. Available from: https://jquery.com/
36. The Broad Institute. SamTools API. Available from: https://samtools.github.io/htsjdk/

37. Highsoft. Highcharts API. Available from: http://www.highcharts.com/products/highcharts. Accessed 19 Dec 2015.
38. IRRI. 3,000 Rice genomes datasets. 2015. Available from: http://oryzasnp-atcg-irri-org.s3-website-ap-southeast-1.amazonaws.com/. Accessed 19 Dec 2015.
39. Oracle. MySQL. 2015. Available from: http://dev.mysql.com/
40. Docker. 2015. Available from: https://www.docker.com/
41. Platform as a Service. Available from: https://en.wikipedia.org/wiki/Paas
42. South Green Bioinformatic Platform. Gigwa code repository. 2015. Available from: https://github.com/SouthGreenPlatform/gigwa
43. Sempere, G; Philippe, F; Dereeper, A; Ruiz, M; Sarah, G; Larmande, P. Supporting information for "Gigwa - Genotype Investigator for Genome Wide Analyses". GigaScience Database. 2016. http://dx.doi.org/10.5524/100199

Wollbrett et al. BMC Bioinformatics 2013, 14:126
http://www.biomedcentral.com/1471-2105/14/126

BMC Bioinformatics

SOFTWARE Open Access

Clever generation of rich SPARQL queries from annotated relational schema: application to Semantic Web Service creation for biological databases

Julien Wollbrett[1,4*], Pierre Larmande[2,4], Frédéric de Lamotte[3] and Manuel Ruiz[1,4*]

Abstract

Background: In recent years, a large amount of "-omics" data have been produced. However, these data are stored in many different species-specific databases that are managed by different institutes and laboratories. Biologists often need to find and assemble data from disparate sources to perform certain analyses. Searching for these data and assembling them is a time-consuming task. The Semantic Web helps to facilitate interoperability across databases. A common approach involves the development of wrapper systems that map a relational database schema onto existing domain ontologies. However, few attempts have been made to automate the creation of such wrappers.

Results: We developed a framework, named BioSemantic, for the creation of Semantic Web Services that are applicable to relational biological databases. This framework makes use of both Semantic Web and Web Services technologies and can be divided into two main parts: *(i)* the generation and semi-automatic annotation of an RDF view; and *(ii)* the automatic generation of SPARQL queries and their integration into Semantic Web Services backbones. We have used our framework to integrate genomic data from different plant databases.

Conclusions: BioSemantic is a framework that was designed to speed integration of relational databases. We present how it can be used to speed the development of Semantic Web Services for existing relational biological databases. Currently, it creates and annotates RDF views that enable the automatic generation of SPARQL queries. Web Services are also created and deployed automatically, and the semantic annotations of our Web Services are added automatically using SAWSDL attributes. BioSemantic is downloadable at http://southgreen.cirad.fr/?q=content/Biosemantic.

Background

Currently, the large amount of plant high-throughput data that have been produced by different laboratories is distributed across many different crop-specific databases. Plant biologists and breeders often need to access several databases to perform tasks such as locating allelic variants for genetic markers in different crop populations and in a given environment or investigating the consequences of a mutation at the transcriptome, proteome, metabolome and phenome levels. The integration of these disparate databases would make complex analyses easier and could also reveal hidden knowledge [1,2].

* Correspondence: julien.wollbrett@cirad.fr; manuel.ruiz@cirad.fr
[1]CIRAD, UMR AGAP, Montpellier F-34398, France
[4]Institut de Biologie Computationnelle, 95 rue de la Galéra, 34095, Montpellier, France
Full list of author information is available at the end of the article

However, biological data integration faces challenges because of syntactic and semantic heterogeneity. In their reviews, Stein LD [3] and Goble C & Stevens R [4] provide a fair criticism of the lack of integrated approaches and provide a similar vision for the future, which is that the Semantic Web (SW) can aid in data integration. According to the W3C, "the SW provides a common framework that allows data to be shared and reused across applications, enterprises, and community boundaries"[a]. The SW currently provides recommendations (RDF [5], SPARQL [6], OWL [7]) for enabling interoperability across databases. Furthermore, major plant databases, such as TAIR [8], Gramene [9], IRIS [10], MaizeGDB [11] and GnpIS [12], annotate their data using ontology terms to link different datasets and to facilitate queries across multiple databases. Guided by life science integration studies

[13,14], annotating data with ontologies promotes the development of ontology-driven integration platforms [15,16].

In parallel, Web Services (WS) are becoming an increasingly popular way of establishing robust remote access to major bioinformatics resources, such as EMBL-EBI, KEGG and NCBI. WS are virtually platform-independent and are easily reusable. Indeed, analysis and data retrieval WSs can be rapidly combined and integrated into complex workflows.

The common use of the SW and WS standards has the promise of achieving integration and interoperability among the currently disparate bioinformatics resources on the Web [17]. There are currently existing efforts to describe Web Services with semantic annotations by using ontologies, such as SSWAP [18], SADI [19] and BioMoby [20]. However, none of these approaches are focused on the automation of business logic [21]. The implementation of new Semantic Web Services (SWS) can be time-consuming and requires the developer to know how to manipulate SW and WS standards and to have expertise on the database schema. To our knowledge, there are currently no ongoing efforts in the context of the automation of SWS creation that are both specific to relational databases and based only on W3C standards.

Our goal is to develop a framework for the creation of SWS for the field of biology by using both SW and WS technologies.

Bio-ontologies result from community reflexions in which each term and each relation are explicitly defined for an application domain. Biological data are annotated with terms from these ontologies, which add a semantic component to them. In BioSemantic, semantics is given by annotation with ontological terms of heterogeneous relational databases schema. These annotations will be used for automatic SWS creation. They will also be used to add semantics to these SWS by annotating their interfaces (input and output).

To make the process of WS development as easy as possible, we have developed a semi-automated framework to accelerate the development of SPARQL queries for relational databases. These queries are automatically added to SWS backbones allowing an easier integration of distributed relational databases. This article focuses on biological relational databases, but because of using only SW and WS standards, BioSemantic can potentially be applied to other science fields.

System and methods

BioSemantic framework overview

The overall architecture of the BioSemantic framework is shown in Figure 1. One advantage of this architecture is that its decoupling takes place in two different steps, which might be achieved by different user profiles. In the first step, the data provider must publish the schema of its relational database. First, the local RDF view of the database schema is automatically created for each relational database to be integrated. Then, the RDF view must be manually annotated by experts with terms from existing bio-ontologies. The RDF views, both created and annotated, are stored in an RDF repository. Once the RDF view is available, the second step is the creation of the SWS. This step is uncoupled from the first step and could be realised by a data consumer without any knowledge of the database schema. The previous semantic annotations of RDF views are used to automatically create SWS containing SPARQL queries and to use the bio-ontological terms as input/output. SWS are then stored in a Semantic Web Services repository, from which they can be easily detected by clients. These clients can use the SWS as wrappers to overstep the heterogeneity of the relational databases.

We will detail below the entire process for generating a BioSemantic SWS, which can be divided into two main parts: *(i)* the generation and semi-automatic annotation of an RDF view (Figure 2) and *(ii)* the automatic generation of the SWS (Figure 3).

Generation and semi-automatic annotation of an RDF view

Relational database-to-RDF mapping

The research in the domain of mapping between databases and ontologies is very active and corresponds to various motivations and approaches [22]. In BioSemantic, we use the mapping as an intermediate layer between the user and the stored data. This layer provides an abstraction of the database and allows the user to query databases without knowledge of the database schema. These characteristics correspond to the motivation known as "data access based on ontology". For that purpose, we found only two tools that strictly use SW standards: Virtuoso [23] and D2RQ [24-26]. We have chosen D2RQ because this tool is open source, easy to use and all of the needed functionalities are free. In addition, some bioinformatics projects have successfully used D2RQ. With D2RQ, we can automatically generate a mapping file that provides an RDF view of the database schema.

RDF view description

The RDF view created by D2RQ can be seen as a mediator of a mediation system. It is used as an interface between the local schema of a database and the global schema defined by bio-ontologies. It is possible to detect all of the heterogeneous RDF views that are annotated with the same ontological term and then retrieve data from corresponding relational databases.

The RDF view generated by D2RQ contains the elements of the database schema: entities, attributes, keys (primary, foreign) and metadata, such as the database driver and host. The data contained in the relational

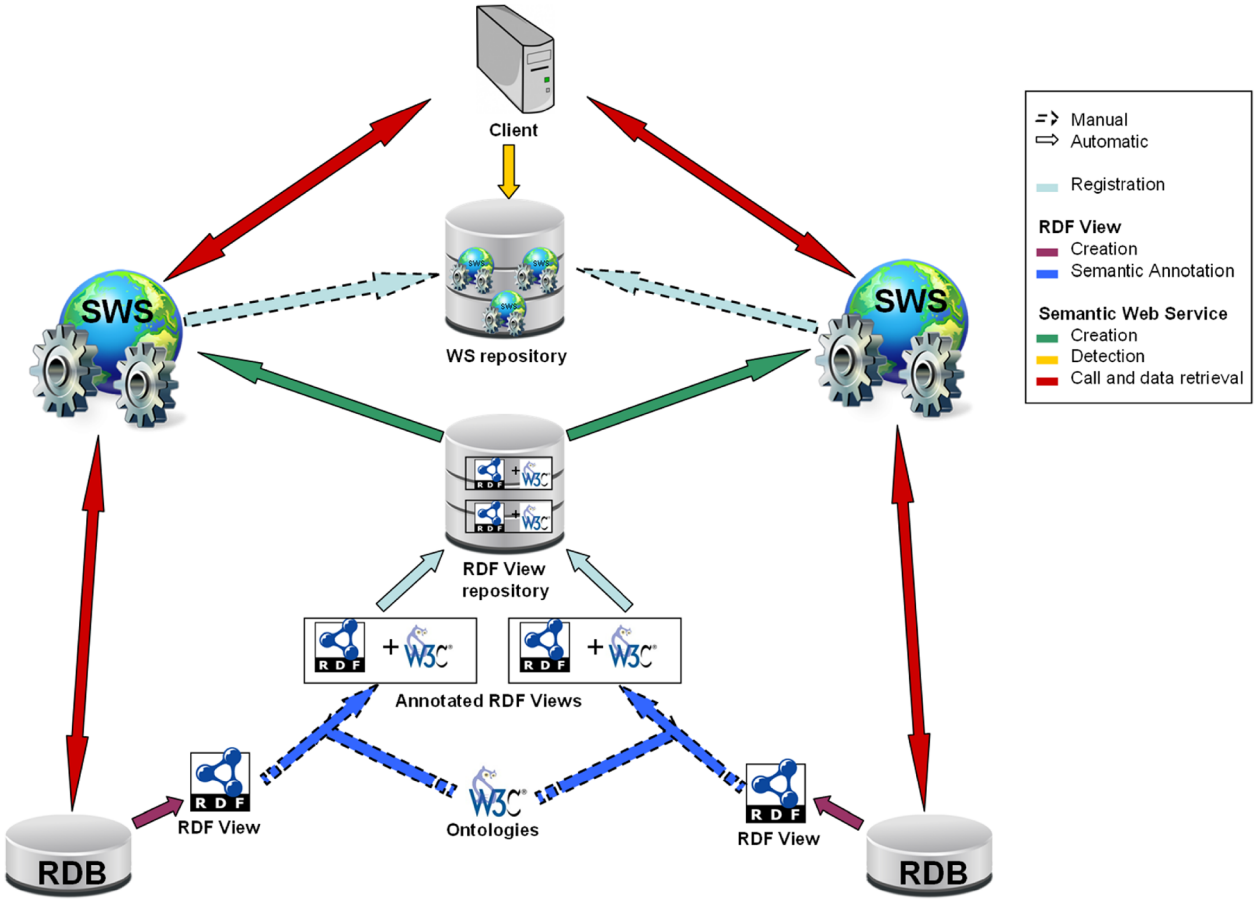


Figure 1 Global architecture of the BioSemantic framework. The first contribution of our work is the automatic creation of an RDF view containing RDF metadata, which is necessary for the automatic creation of Semantic Web Services. The second contribution is the automatic creation and deployment of Semantic Web Services.

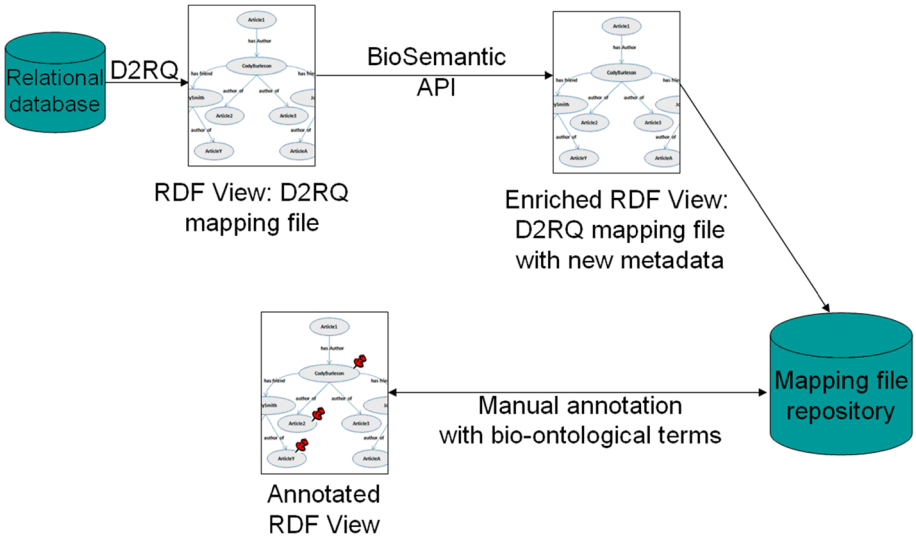


Figure 2 Generation and semi-automatic annotation of the RDF view. D2RQ creates the D2RQ mapping file, and our BioSemantic API automatically adds new metadata about the database schema. Finally, the mapping file is stored in a repository. Annotation with bio-ontological terms is performed manually by an expert.

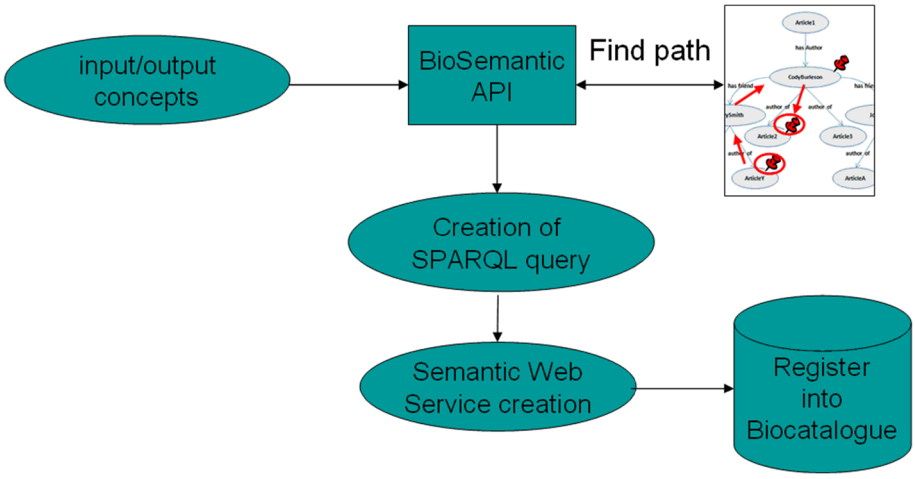


Figure 3 Automatic generation of Semantic Web Services. The Web Services developer selects the bio-ontological terms to be used as input/output. All of the mapping files, which are stored in the mapping file repository, are automatically parsed to find a path linking the input and output ontological terms. If such a path is found, it is used to create a SPARQL query. The query is integrated into a semantic Web Service that is then registered in a Web Service registry, such as BioCatalogue.

databases are not included in the RDF view. Instances are retrieved directly from the databases. D2RQ API uses metadata from the RDF views to connect to the databases and to retrieve instances from them. The RDF view is queried with a SPARQL query; then, the D2RQ API transforms this query into an equivalent SQL query. Thus, there is no problem with keeping data up-to-date because the data are not physically exported.

In the RDF view, the database schema is represented by a graph. Each node corresponds to an entity or attribute in the database, and each edge defines a relationship between two nodes. In RDF format, namespaces are used to uniquely identify each node. Namespaces provide a prefix for each node name. For example, the *map:marker* node (Figure 4) indicates the "marker" concept from the "map" vocabulary used by D2RQ to uniquely identify one RDF view and to map relational elements to the RDF view.

Automatic semantic enrichment of the RDF view with BioSemantic

The BioSemantic API automatically detects specific information related to the relational database schema and translates it into new properties that can be integrated into the RDF view. These metadata are then used for

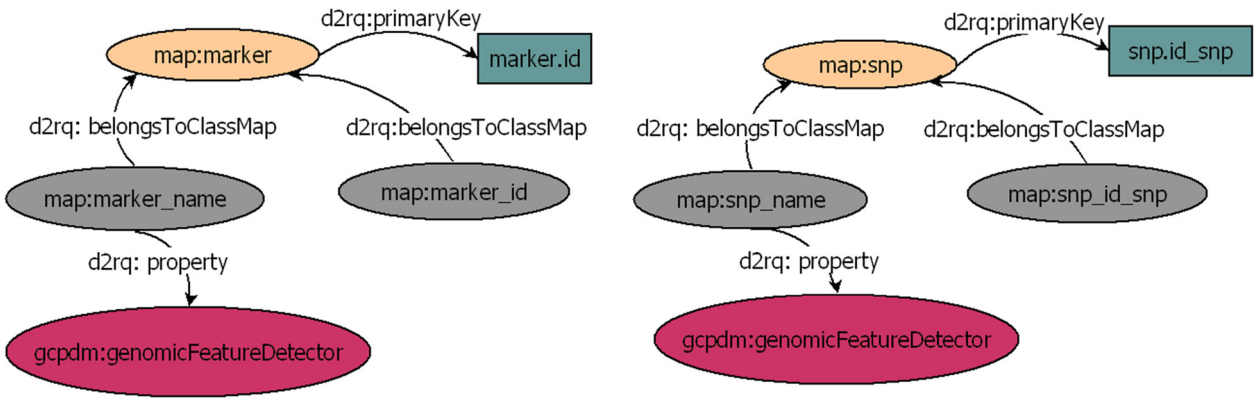


Figure 4 Graph-based representation of annotated RDF views. Each graph is the RDF representation of some part of a relational database. The *d2rq:belongsToClassMap* property links a column to a table. The *d2rq:primaryKey* property defines the primary key of a table. The *d2rq:property* property links a node to a semantic annotation. The columns *marker_name*, from the table *marker*, and *snp_name*, from the table *snp*, are both annotated with the same term: *genomicFeatureDetector* from the GCP domain model ontology [27].

SPARQL query generation. This step can be seen as a semantic enrichment of the RDF view.

1. *Association tables*

For this purpose, we have developed an algorithm that detects association tables.

$$
\begin{aligned}
&pk = \text{primary key of } R\\
&fk = \text{foreign keys of } R\\
&\text{if } ((\forall u \in R)(u \in fk \Rightarrow u \in pk)) \{\\
&\qquad \text{if } ((\forall u \in R)(u \in pk \Rightarrow u \in fk)) \{\\
&\qquad\qquad R \text{ is an association table}\\
&\qquad \}\\
&\}
\end{aligned}
$$

2. *Arity*

We can also detect the arity of association tables, i.e., the number of foreign keys that they possess. The algorithm labels association tables in the RDF view with the *dr:associatedTo* property and indicates the arity with the *dr:arity* property (Figure 5).

3. *Inheritance, aggregation and composition*

There are many ways to transform inheritance relationships from an object-oriented conceptual model to a relational model [28]. For our algorithm, we detect relationships that result from the transformation of each class in an inheritance hierarchy into a table. We also detect tables that result from aggregation or composition relationships by using the identifying algorithm from [29]. We label these relationships in the RDF view with the *rdf:subClassOf* property (Figure 5).

Manual annotation with bio-ontological terms

The D2RQ language allows elements of the mapping file to be annotated with bio-ontological terms, which can be interpreted as semantic flags. Such flags can be used directly to query the relational database without any prior knowledge of the database schema or can be used to locate corresponding elements across databases

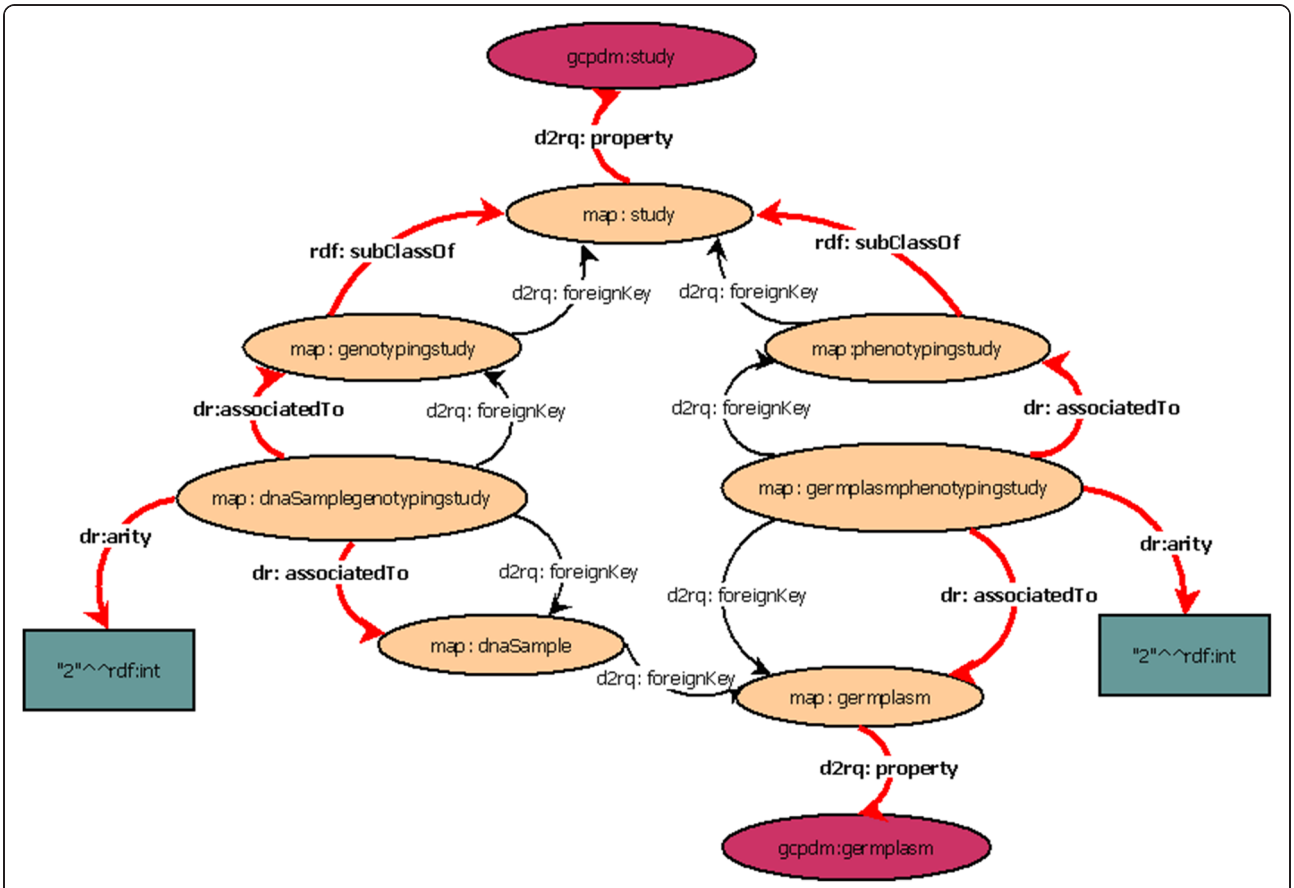


Figure 5 Classification of the database table relationships. Each light node represents a table of the relational database. Here, we only show the tables, and the columns are not represented. The dark nodes represent the semantic annotations. Each edge represents a property that is shared between 2 nodes. The new properties added by our method, *dr:associatedTo*, *dr:arity* and *rdf:subClassOf*, are indicated in bold.

(Figure 4). The annotation of the RDF view is performed manually by adding triples to the RDF view using a text editor and must be conducted by an expert familiar with both the database and the bio-ontology. In the plant biology domain, some ontologies are implemented in OBO format and do not provide URLs, in contrast to OWL ontologies. For this reason, the terms used to annotate the RDF view can be explained as URIs that do not resolve. Nevertheless, according to W3C standard it is recommended to use URLs that resolve.

Automatic generation of the Semantic Web Service

Semantic annotations are used to select the inputs and outputs of a query. We can find a path in one RDF view by linking the inputs to the outputs. If such a path is found in the RDF view, then it is used to create a SPARQL query. To automate the creation of SPARQL queries, we implement an algorithm that is a single-pair variant of the shortest-path algorithm. Given an input graph, a source node and a destination node, the algorithm returns a path linking the two nodes through the graph. We add conditions to our shortest-path algorithm according to the types of relationships between the nodes, which can be either of the following: *(i)* relationships that correspond to association tables; or *(ii)* relationships that result from inheritance, aggregation, or composition in an object-oriented conceptual model. These conditions correspond to the metadata that is added to the RDF view during the automatic semantic enrichment step that is taken by the BioSemantic API.

Shortest-path algorithm with conditions

We parse the RDF view as though it were a graph, to find the shortest path linking two bio-ontological terms. These terms correspond to those selected as input and output for our WS.

We use a shortest-path detection approach based on the Dijkstra algorithm [30]. We add conditions to the weight path costs according to the properties classified in the previous step. In the weighting, we favour paths that correspond to binary associations. For the shortest paths that correspond to the *rdf:subClassOf* property (inheritance, aggregation or composition), we aggregate the different paths found. For example, in Figure 5, the *rdf:subClassOf* property allows a study to be considered a *genotypingStudy* or a *phenotypingStudy*. The data recorded in these two tables are complementary and are non-redundant. Indeed, the path linking *gcpdm:study* to *gcpdm:germplasm* is the combination of both paths:

Path 1: *map:study- > map:genotypingstudy- > map:dnasamplegenotypingstudy*

- > map:dnasample- > map:germplasm

Path 2: *map:study- > map:phenotypingstudy- > map:germplasmphenotypingstudy*

- > map:germplasm

These paths are not stored; instead, they are dynamically detected and are used to create a SPARQL query.

Generation of SPARQL queries

The detected path contains all of the information that is required for the automatic creation of a SPARQL query. For a given set of input/output bio-ontological terms and a given RDF view, only one SPARQL query can be created. The query below corresponds to the link between *gcpdm:study* and *gcpdm:germplasm*. *SELECT DISTINCT ?study_name ?germplasm_name WHERE {*

```
?study_id gcpdm:study ?study_name.
FILTER regex(?study_name,"^name_of_the_study$").
{
?genotypestudy_id vocab:genotypingstudy_id_study ?
study_id.
?key vocab:
dnasamplegenotypingstudy_id_genotypingstudy ?
genotypestudy_id.
?key vocab:dnasamplegenotypingstudy_id_dnasample ?
dnasample.
?dnasample vocab:dnasample_id_germplasm ?
germplasm_id.                    Path 1
?germplasm_id gcpdm:germplasm ?germplasm_name.
}
UNION {
?phenotypestudy_id vocab:phenotypingstudy_id_study ?
study_id.
?key vocab:
germplasmphenotypingstudy_id_phenotypingstudy ?
phenotypestudy_id.
?key vocab:germplasmphenotypingstudy_id_germplasm ?
germplasm_id.   Path 2
?germplasm_id gcpdm:germplasm ?germplasm_name.
}
}
```

The first line of the query defines the attributes that correspond to the input and output of the WS. The third line is always a FILTER condition. This filter applies to the input attribute, which can be a literal or a regular expression. In our example, it is possible to retrieve the names of the germplasms that are used in a study by using names that begin with A and the regular expression "A.*".

Automatic creation of the Semantic Web Service

The SPARQL query is automatically integrated into a WS template. The WS is annotated with the bio-ontological terms previously selected as input and output for the query. According to the recommendations of the EMBRACE project [31] and the W3C, we use the

Semantic Annotations for WSDL (SAWSDL) [32] to add semantic annotations to the WSDL (Web Services Description Language) components. The use of SAWSDL offers three main advantages: *(i)* it is compatible with the WSDL standard; *(ii)* it is lighter than other computing standards (i.e., WSMO (Web Service Modeling Ontology) and WSDL-S (Web Service Semantics)); and *(iii)* it is recommended by the W3C. Indeed, the input/output of our SWS are annotated using the *sawsdl:modelReference* attribute, specifying the association between an WSDL component and a bio-ontological term (Figure 6).

One SWS is created for each detected SPARQL query. All of the SWS annotated with the same input/output concepts can be easily detected and used for data integration. After the SWS is created, it can be registered in Web Service registries, such as BioCatalogue [33].

Implementation

Our method is implemented in Java. The RDF views are created using the d2rq 0.7 library, and the RDF files are parsed using the Jena 2.5.7 library. The SWS are automatically deployed on a Tomcat 6.0 server using Axis2.

Results

Use case

We have created a use case integrating *Oryza sativa* (rice) data from distributed relational databases: Gramene [9], TropGene [34] and Ensembl [35]. Both the Gramene and TropGene databases have QTL data associated with traits, and these traits can be associated with concepts from the Trait Ontology. We wanted to compare the rice QTLs from the two resources, Gramene and TropGene, and to extract related genomic annotations from the Ensembl rice module.

We first used BioSemantic to create the SWS. We then used Taverna [4] to create a workflow by connecting BioSemantic SWS with external public WS. In this manner, we could verify the compatibility of BioSemantic SWS with standard WSDL WS. To increase the speed of querying over huge tables, we used a local copy of the Markers tables of Gramene; however, our example performed adequately using a remote access to the Gramene public database.

All automatic steps can be performed directly on the BioSemantic Web user interface (Figure 7).

Steps involving SWS creation and using the BioSemantic Web user interface

A simple form must be completed to configure database access and to automatically create RDF views for the TropGene and Gramene databases (Figure 7). The RDF views can then be downloaded to perform semantic annotations. In our example, we annotated RDF views using one concept from the EDAM ontology [31]. The elements of the RDF views were annotated with the same ontological concept, known as edam:1093. For readability, we choose to represent this concept by its name edam:sequence_accession in our example. This annotation is added to triples corresponding to the `marker`.`name` column of the RDF View of TropGene. The annotation is represented below in bold type.

map:marker_name a d2rq:PropertyBridge;

d2rq:column "marker.name";
d2rq:property edam:sequence_accession;
d2rq:belongsToClassMap map:marker;

An annotation with the same term is added to triples corresponding to the `marker`.marker_acc` column of Gramene. The annotation is represented below in bold type.

map:marker_marker_acc a d2rq:PropertyBridge;

d2rq:column "marker.marker_acc";
d2rq:property edam:sequence_accession;
d2rq:belongsToClassMap map:marker;

The same ontological term is then used to annotate different database schemas. The BioSemantic Web interface allows users to upload the annotated RDF views to visualise the list of available RDF views in the repository, to download one of the views in order to view/add/modify

```
<xs:element name="method">
 <xs:complexType>
  <xs:sequence>
   <xs:element minOccurs="0" name="input" nillable="true"  type="xs:string"
   sawsdl:modelReference="http://gcpdomainmodel.org/GCPDM#GCP_GenotypeStudy"/>
  </xs:sequence>
 </xs:complexType>
</xs:element>
```

Figure 6 SAWSDL annotation. The semantic annotation is represented in bold and tags the input of our Semantic Web Service with the GCP_GenotypeStudy term from the GCP domain model ontology.

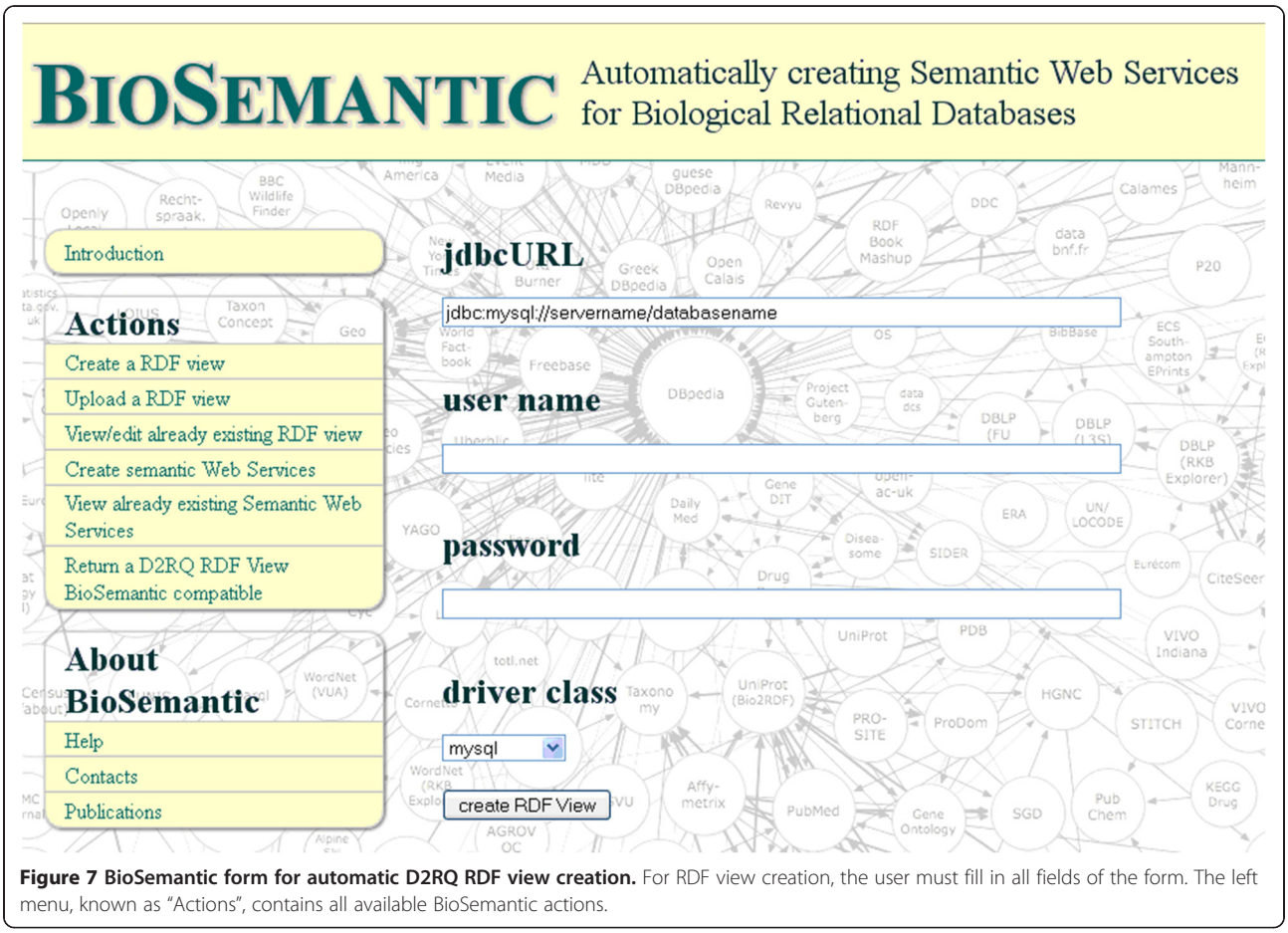


Figure 7 BioSemantic form for automatic D2RQ RDF view creation. For RDF view creation, the user must fill in all fields of the form. The left menu, known as "Actions", contains all available BioSemantic actions.

annotations and to visualise the list of ontology and concept terms currently used into the RDF repository. This interface also allows users to automatically add BioSemantic annotations to a pre-existing D2RQ RDF view. Some projects use D2RQ, which means that some RDF views are currently annotated with domain ontologies. This functionality allows users to return these RDF views to BioSemantic compatibility without manual steps.

After selection of the input/output bio-ontological terms (Figure 8), the BioSemantic application displays the list of RDF views containing these annotations (the red box in Figure 9). The checkbox before the name of an RDF view allows the user to select the SWS that he would like to create. By clicking on the radio button, the corresponding SPARQL query is displayed. It is then possible to validate the automatically generated query or to modify it (e.g., add more filters). A simple click on a button then creates SWS files and deploys them.

Workflow creation

BioSemantic SWS can be obtained directly with their WSDL localisation. In this use case, we chose to compose SWS as a workflow in Taverna. Taverna makes it possible to easily create a workflow, to visualise the progress of the running workflow and to save the workflow for the purpose of sharing it.

Our workflow (Figure 10) contains 7 BioSemantic SWS (green boxes). Yellow and purple boxes correspond to bricks that transform the inputs/outputs of the SWS and then allow for composition. For a given trait and QTL maximum size, BioSemantic SWS retrieve the Gramene and TropGene accession numbers of the QTLs along with their mapping positions in *Oryza sativa*. We have created a Beanshell Taverna brick (orange box) to retrieve the Gramene and TropGene QTLs that are mapped in the same genomic region. Two other bricks allow for the compatibility of the BioSemantic SWS with the Ensembl BioMart WSs. Indeed, we added Ensembl BioMart WS (blue boxes) to retrieve genes that are present in the mapping genomic interval of a given QTL. The yellow and purple boxes are shims that are added in Taverna. The purples boxes allow Taverna to manipulate BioSemantic SWS input. They are created automatically by Taverna. The yellow boxes are XPath expressions that allow Taverna to

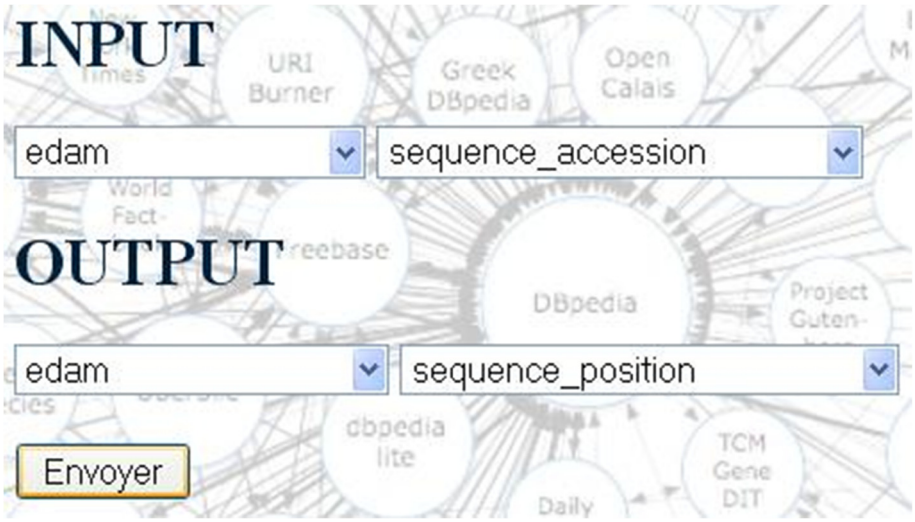


Figure 8 BioSemantic form for input/output concept selection. These concepts will be used to detect a path and to annotate the input/output of the SWS. The user can only select the prefix and concepts used to annotate a previously registered RDF view.

be compatible with BioSemantic SWS output. All of the yellow boxes are identical, and their creation is fast. However, the presence of these shims does not allow automation of BioSemantic SWS compositions.

In brief, our workflow retrieves the following rice information from TropGene, Gramene and Ensembl:

- Accession number of the QTLs associated with a given trait,
- Pair-based position of the mapping of these QTLs,
- All of the genes in the mapping interval of a given QTL, and
- QTLs with a common mapping position between TropGene and Gramene.

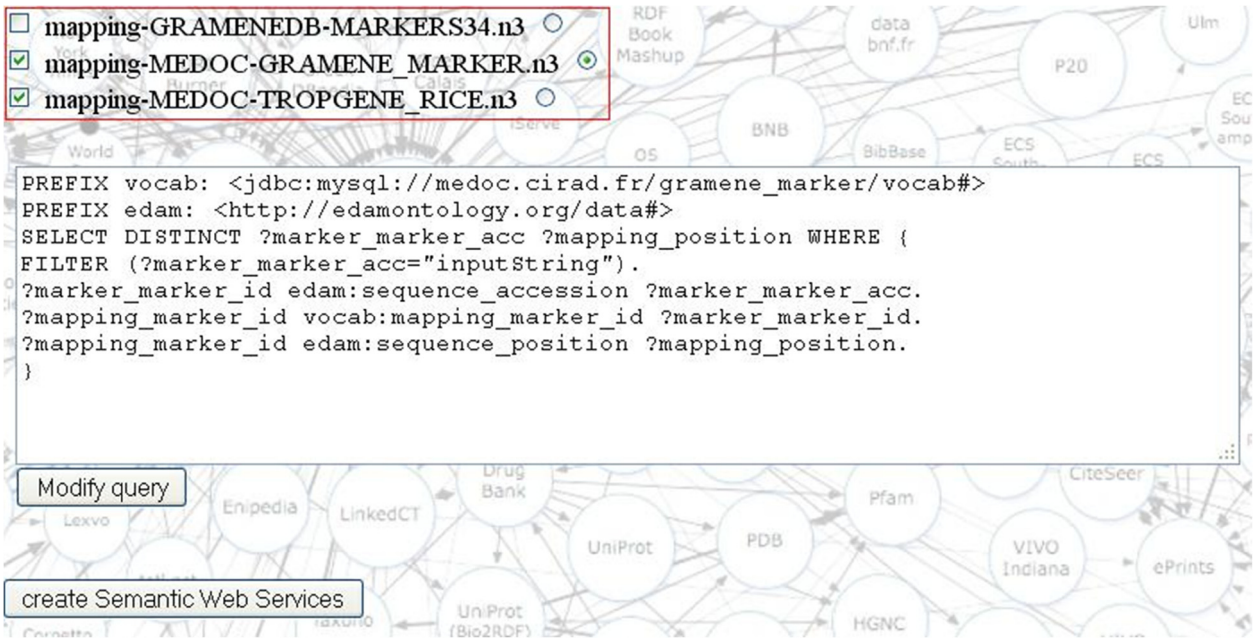


Figure 9 BioSemantic form for RDF view selection and query visualisation/edition. The red rectangle contains the name of the RDF views annotated with both input/output concepts. The checkbox before each name allows for the selection of a view for SWS creation. The radio button after the name of an RDF view allows for query visualisation/edition. When all desired RDF views are selected, a simple click creates the SWS.

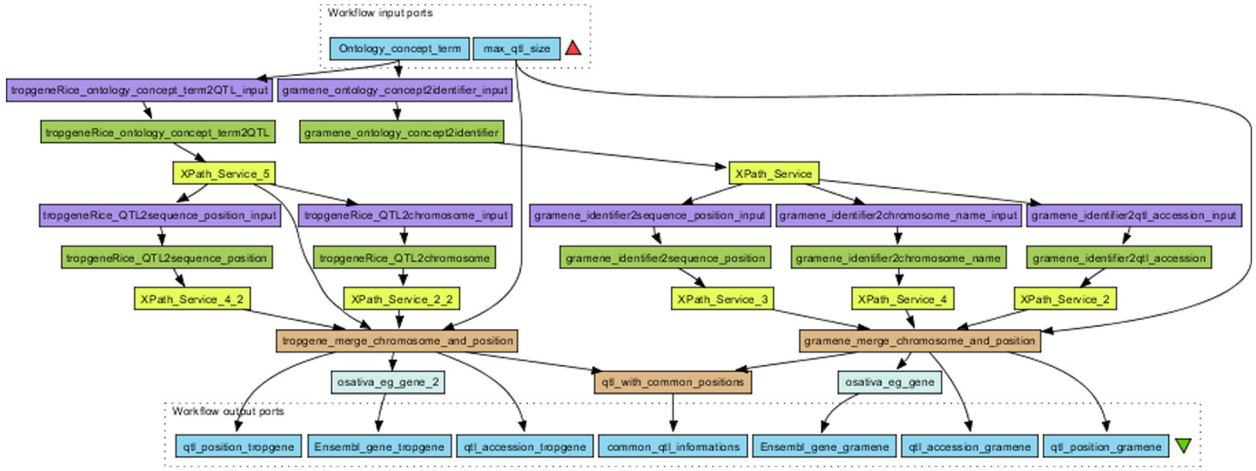


Figure 10 Use case workflow created with Taverna 2. The workflow contains BioSemantic SWS querying for both the TropGene and Gramene databases for QTL information retrieval and also contains BioMart WS querying EnSembl for gene information retrieval. This workflow also detects QTLs from TropGene and Gramene when both are annotated with the same TO term and have the same mapping position.

– This workflow can be downloaded in my Experiment [36].

Available Semantic Web Services

We developed other SWS for our own databases, including TropGene and OryGenesDB, a database of functional rice genomics data [37], as well as from external databases such as Gramene and SINGER [38], a multi-crop germplasm database. We annotated the database schemas with concepts from the Crop Ontology [39], the GCP Domain Model [15], the Sequence Ontology [40] and the EDAM ontology [31]. Some of these generated WS are available in the BioCatalogue (Figures 11 and 12).

Benchmarks

With regard to automatically generated SPARQL queries, we are aware that, in some cases, there are multiple possible paths, each of which can be semantically valid depending on the query semantics. Our system identifies the "best" shortest-path with conditions favouring binary table associations and combines the paths corresponding to inheritance, aggregation and composition. However, a manual validation test for the automatically generated SWS is still recommended. Indeed, the SWS that we generated and tested were all validated by the database managers and/or users. Regardless of the validation ability, the main benefit of our platform is that it enables the rapid creation of new and easily detectable SWS.

SPARQL query generation

We have tested the speed of SPARQL query generation with two different biological databases: *(i)* TropGene, a relational database that contains 90 tables and 15 million records; and *(ii)* OryGenesDB, which contains 11 tables and 22 million records. Although SPARQL query generation is only performed during the first step of Web SWS generation and not during the SWS execution, we also measured the time required for this step. This time depends on the database schema but also strongly depends on the presence of inheritance relationships (Table 1). In this table, when inheritance relationships are present, we include the lengths of the paths to be aggregated. The creation of a query without inheritance relationships takes less than 2 seconds. However, creating a query using the same database schema with 4 inheritance relationships takes 15 seconds. In general, complex SPARQL queries can be created in a matter of seconds.

Port: GetTropgeneMarkerSPARQL
Location: http://gohelle.cirad.fr:8080/testWS/services/GetTropgeneMarkerSPARQL
Protocol: http://schemas.xmlsoap.org/soap/http
Default Style: document

Figure 11 General information about the GetTropgeneMarkerSPARQL Web Service registered in the BioCatalogue.

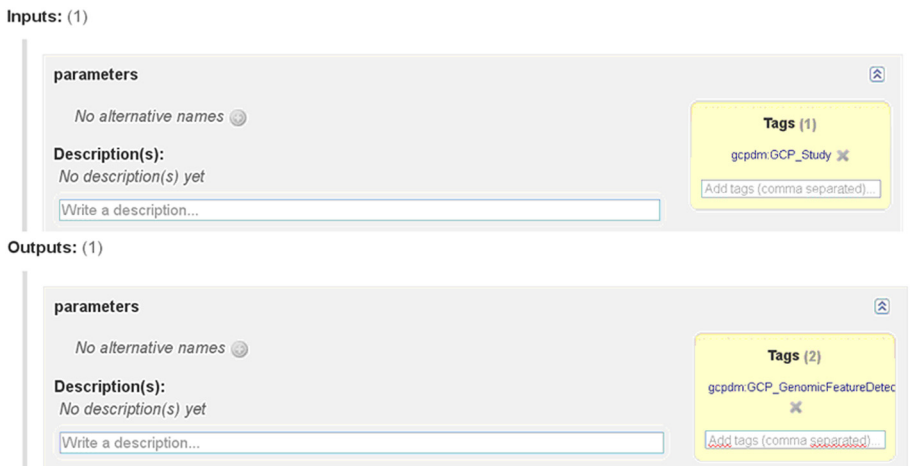


Figure 12 BioCatalogue input/output annotations for the GetTropgeneMarkerSPARQL semantic Web Service. Our bio-ontological terms correspond to the input/output tags in the BioCatalogue.

SPARQL query execution

The time required for query execution varies significantly for different databases and strongly depends on the number of records to be retrieved. Table 2 compares the time required for SQL query execution and SPARQL query execution. We did not compare with the time required for the querying RDF dump of a relational database because some databases can contain more than 100 million tuples. The RDF dump will then contain more than 100 million triples, and triplestore query performances decrease with the number of triples. When the triplestore contains more than 100 million triples, SPARQL to SQL approaches are fastest [41]. The time required for SQL query execution was measured in Eclipse using the *java.sql* library. The time required for SPARQL query execution was measured using the AJAX-based SPARQL Explorer tool of the D2R Server.

The SPARQL approach takes approximately 3-4 times longer to access data than a direct SQL query, but users can still retrieve more than 5000 results in a few seconds. The time required to display the SPARQL results in the AJAX-based SPARQL Explorer accounts, in part, for the differences in performance. Most of the overhead, however, comes from the transformation of SPARQL queries into SQL queries, which is performed using the D2RQ engine.

Table 1 Estimating the time required for SPARQL query creation

Number of tables	Inheritance relationship	Length of the path	Time (seconds, ± 0.1)
11	no	2 nodes	1.2
90	no	4 nodes	2.0
90	yes	4-3-2-6 nodes	14.6

Semantic Web Services execution

Table 3 compares the time required for SWS execution using manually created SQL WS and our automatically generated SPARQL SWS. These Web Services query the TropGene database. Although manually created WS are faster than our automatically created SWS, the difference is not dramatic enough to affect the usability of our SWS.

Validation of the SPARQL query results

We compared the data retrieval resulting from the three approaches (i.e., the Dijkstra algorithm, BioSemantic and a human SQL query builder) (Table 4). We refer to a human SQL query as a query that is manually written by an expert with good knowledge of the database schema. A first general observation demonstrates that the number of results is identical for BioSemantic queries and the manual SQL queries. BioSemantic globally retrieves more results than the Dijkstra algorithm. The gap for Query1 is explained because of the inheritance relationships missed by the Dijkstra algorithm. Indeed, in that case, BioSemantic detects these relationships and regroups the subdivided paths into the final query. Furthermore, BioSemantic preferentially selects binary association tables that promote more data retrieval. Both Query2 and Query3 correspond to a short path without inheritance but with several paths having the same node numbers. In that case, weighting the BioSemantic path favours binary associations, whereas the Dijkstra algorithm chooses the first detected path having a minimum node number. For Query2, BioSemantic favours the detection of a more pertinent path, whereas the same paths are detected for Query3. For Query4, no equivalent path guides to the same results; in other words, both algorithms select the same path. In each case, we manually verified that the retrieved data were identical.

Table 2 Comparison of the time required for SQL and SPARQL query execution

Number of tables	Inheritance relationship	Length of the path	Number of results	SQL query (seconds, ± 0.1)	SPARQL query in D2R Server (seconds, ± 0.1)
90	no	4	860	0.4	1.4
90	no	4	1456	0.4	1.4
90	no	2	2055	0.8	2.3
90	yes	4-3-2-6	8071	1.1	4.2
90	no	3	12302	2.3	4.8

Comparison with other SWS platforms

We compared BioSemantic with other SWS platforms, such as BioMoby [20], SADI [19] and SSWAP [18] (Table 5). BioMoby adds semantic components to WSs by using an XML datatype ontology developed by WS developers. SSWAP is based on a five-class ontology allowing the definition of Web resources, inputs and outputs of the SWS, data structures and data providers. SADI is a set of fully standard-compliant SWS design patterns that simplify their publication. A SADI plugin has been developed. This plugin helps users to discover SADI SWS and to automatically compose them in workflows.

In this comparison, we focused on the ability to create and use SWS because the other SWS approaches are not placed in the context of the automated creation of wrappers for relational databases.

We compared seven criteria: i) the exclusive use of SW standards; ii) the types of input and output annotation for SWS; iii) the compliance with SOAP/WSDL; iv) the constraint for clients to be platform specific; v) the ability of the platform to perform reasoning; vi) the degree of automation in the creation and deployment of SWS; and vii) the degree of automation of the query building.

All of the compared approaches use SW standards except for BioMoby, in which semantics come from the data type stored in an XML tree. In terms of output, SADI and SSWAP are based on OWL, and both developed their own SWS API to exploit OWL's reasoning capabilities. BioSemantic uses the standard SAWSDL to semantically annotate the WSDL files.

BioMoby, SADI and BioSemantic are compliant with SOAP/WSDL protocols. Some of the approaches are platform specific (i.e., SSWAP and BioMoby), meaning that they require their own environment to process SWS. For example, SSWAP gains in speed and lightness but loses in genericity. BioMoby develops its own data type definition, allowing for an easy choreography of services, but requires clients to be compliant with the API. BioSemantic and SADI use standard clients to call their SWS.

In terms of reasoning abilities, SADI and SSWAP exploit OWL with semantic reasoners to highlight some relationships between classes. On the other hand, BioMoby exploits the taxonomic properties of XML to infer relationships between data types; however, BioMoby is less expressive than OWL. BioSemantic comes without reasoning capabilities. Initially, this task was to be performed by the SWS catalogue (i.e., BioCatalogue), but this function is not yet available.

The last two criteria define the degree of automation of these approaches. BioMoby and SADI allow for the creation and deployment of SWS skeletons without including core methods. BioSemantic is the only API that processes query creation. This automation is allowed by decoupling annotated RDF view creation and SWS creation. However, this automatic creation of SWS is still dependent on the manual RDF view annotation step performed by the data provider.

Table 3 Comparison of the time required for Web Service execution using the SQL Web Services and automatically generated using the SPARQL Web Services

Query	Number of results	SQL Web Services (seconds, ± 0.1)	SPARQL Web Services (seconds, ± 0.1)
retrieves genotyping studies	7	0.2	1.0
retrieves germplasms for selected studies	860	0.4	1.0
retrieves markers for selected studies	1456	0.4	1.0

Discussion

Semantic limitations

OBO ontologies

The development of an ontology is a long community-based task in which participants decide on a consensus basis about term definitions and relationships between

Table 4 Comparing the number of retrieved data from the three approaches: Dijkstra algorithm, BioSemantic and human SQL query builder

	Inheritance	Equivalent paths	Dijkstra	BioSemantic	Manual SQL
Query 1	yes	no	1595	7212	7212
Query 2	no	yes	0	12302	12302
Query 3	no	yes	197	197	197
Query 4	no	no	2055	2055	2055

Table 5 Comparison with other SWS platforms

	Semantic Web Standard	Annotations	WSDL compliant	Platform specific	Reasoner	Creation/ deployment	Query creation
BioMoby	no	XML	yes	yes	no	semi-automatic	manual
SSWAP	yes	OWL	no	yes	yes	manual	manual
SADI	yes	OWL	yes	no	yes	semi-automatic	manual
BioSemantic	yes	SAWSDL	yes	no	no	automatic	automatic

those terms. Currently, a large number of bio-ontologies exist and cover a large spectrum of biological domains.

Most of these ontologies are not developed in an OWL format; instead, they are in an OBO format, which follows the OBO Foundry principles [42], such as unique URI or formatted term/concept names.

Regarding the amount of work that is necessary to create an ontology, we decided to allow the annotation of RDF view using terms from OBO ontologies. However, that strategy could raise problems, such as the possible lack of a URL that could resolve these ontologies. However, even if OBO Foundry principles only recommend using unique URIs, a lot of already existing OBO ontologies are associated to URLs. Furthermore, if OBO ontologies do not use URLs that currently resolve, it is still possible to register them with online tools such as BioPortal or Ontology Lookup Service (OLS). In our case, we deployed an instance of OLS allowing publishing ontologies on the Web.

In our approach, the major limit from OBO ontologies comes from the low number of classes possessing restrictions along with the low number of different properties used (e.g., BioPortal notes that 8 properties are used in the GO, which possesses more than 38000 classes). Therefore, using those ontologies has a strong impact on BioSemantic by significantly limiting its semantic component.

Manual SWS composition

The SWS BioSemantic composition requires the development of shims. This requirement is a limit to the workflow creation that could be overtaken by creating a Taverna plugin or by making the BioSemantic framework compatible with SADI. Moreover, SADI already possesses a Taverna plugin. Furthermore, that compatibility could take advantage of a stronger semantic without being platform specific.

No use of existing framework

In BioSemantic, we choose to not reuse already existing frameworks such as BioMoby, SSWAP or SADI. Indeed, the purpose of these frameworks is to better organise semantic components, whereas the main purpose of BioSemantic is to separate the steps of publishing relational schema and the creation of SWS and then to automate the step of SWS creation.

During our work, we did not focus on making our approach compatible with an already existing framework. The main reason was that we did not want to be affected by the technical or compatibility limits of other WS or by the success of our approach depending on a specific framework. However, SSWAP and SADI are based on OWL, which allows the creation of SWS with stronger semantics than BioSemantic. Using BioSemantic in those frameworks could increase widely the semantic component of SWS created by BioSemantic and therefore automate their composition.

Differences between semantic and data type

The use of bio-ontology terms to annotate input/output allows for easier detection of our SWS by searching services with a standard vocabulary.

Annotations are composed of adding a semantic flag on a component of a database schema, which requires choosing which component of a schema will be annotated.

That step is performed manually, and does not guarantee that the same annotation will be associated with similar data. For example, we used the term gcpdm:study to annotate the name of a study because the only identifier of a study existing in the TropGene database is an auto increment with no scientific sense. If another curator uses the same term to annotate an identifier, the data returned by the two different services would not be comparable even if the two services return information on a genotyping study. That limit prevents the automatic composition of our services into workflows.

Shortest path algorithm

One input and one output

The major limit of our query comes from the restriction to a single input concept and a single output concept. That restriction is because of the shortest path algorithm, which allows only the joining of a node of a graph to another node. That restriction implies that we create a query coming from a linear path in the graph that represents the database schema.

It would be interesting to modify our algorithm to find a path that links a number n of input nodes in our future query to n output nodes.

Retrieve one path

Currently, BioSemantic allows automatic query creations based on our shortest path algorithm. We plan to allow a user to choose between different paths. The visualisation of these paths, in which nodes correspond to database table names, will aid in user selection.

Self join detection

Furthermore, BioSemantic does not allow the creation of queries annotated with the same input and output concept as a consequence of using the Dijkstra algorithm. This functionality would be very interesting for orthologous or synonym detection for example.

We plan to implement a simple algorithm allowing the detection of all self joins that correspond with a given table. In fact, if a table has several self joins, then the path length found for each of them will be identical. For this reason, both the path visualisation in the graphical interface for the query creation and the algorithm of self join detection will overtake this limitation, allowing the user to select the name of the wanted association table, to link one table to itself and then to create the wanted query.

Manual RDF view annotation

Future developments will concern the semantic annotation of RDF views, which is the only manual task in BioSemantic. This task could be a time-consuming task for database annotators if the database schema is large. Constraints for annotators arise many because they must be experts on database schemas and the ontology terms and must also manipulate RDF and D2RQ. We believe that this limitation could be partially overcome by creating a user interface for the annotation of RDF Views.

Our solution opens new perspectives for the development of SWS. However, we are still interested in adding more functionality, such as the automatic generation of links between database schemas and existing ontologies. We are currently exploring the use of automatic schema-matching tools developed in the context of the WebSmatch platform [43].

Performance problems with FILTER

Input variables of our services can be regular expressions or literal expressions. Those variables are detected when WS is used and will lead to the use of a different SPARQL FILTER. A regular expression used in a WS input could raise query problems, for example, it could create memory errors when querying tables that contain hundreds of thousands of tuples.

Conclusions

BioSemantic is a framework that is designed to speed the development of Semantic Web Services for existing relational biological databases. This framework has the specific capability of separating the publishing step of the relational schema from the SWS creation. Data consumers can then create Semantic Web Services without knowledge of the resource schema. Currently, it automatically creates and semi-automatically annotates RDF views that enable the automatic generation of SPARQL queries. These queries are created by the following steps: *(i)* the selection of input and output ontological terms using a Web interface that is available in the BioSemantic API; *(ii)* the automatic detection of a path linking inputs to outputs; and *(iii)* the use of the path to automatically generate a SPARQL query. Semantic Web Services are also automatically created and deployed.

Availability and requirements

- **Project name:** BioSemantic
- **Project home page:** http://southgreen.cirad.fr/?q=content/Biosemantic
- **Operating system(s):** Platform independent
- **Programming language:** java
- **Restrictions on use by non-academics:** no limitations

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

JW developed and tested the Java code. All of the authors contributed to the design of the software architecture and the development of the appropriate methods. All of the authors read and approved the final version of the manuscript.

Acknowledgements

We would like to acknowledge Isabelle Mougenot and Guilhem Sempere for their assistance.
This work was supported by Région Languedoc-Roussillon and CIRAD.

Author details

[1]CIRAD, UMR AGAP, Montpellier F-34398, France. [2]IRD, UMR DIADE, Montpellier, France. [3]INRA, UMR AGAP, Montpellier F-34398, France. [4]Institut de Biologie Computationnelle, 95 rue de la Galéra, 34095, Montpellier, France.

Received: 23 August 2012 Accepted: 25 March 2013
Published: 15 April 2013

References

1. Tsesmetzis N, Couchman M, Higgins J, Smith A, Doonan JH, Seifert GJ, Schmidt EE, Vastrik I, Birney E, Wu G, D'Eustachio P, Stein LD, Morris RJ, Bevan MW, Walsh SV: **Arabidopsis reactome: a foundation knowledgebase for plant systems biology.** *Plant Cell* 2008, **20:**1426–1436.
2. Lysenko A, Hindle MM, Taubert J, Saqi M, Rawlings CJ: **Data integration for plant genomics—exemplars from the integration of Arabidopsis thaliana databases.** *Brief Bioinform* 2009, **10:**676–693.
3. Stein LD: **Towards a cyberinfrastructure for the biological sciences: progress, visions and challenges.** *Nat Rev Genet* 2008, **9:**678–688.
4. Goble C, Stevens R: **State of the nation in data integration for bioinformatics.** *J Biomed Inform* 2008, **41:**687–693.
5. *RDF/XML Syntax Specification (Revised).* http://www.w3.org/TR/REC-rdf-syntax/.
6. *SPARQL Query Language for RDF.* http://www.w3.org/TR/rdf-sparql-query/.
7. *OWL Web Ontology Language Overview.* http://www.w3.org/TR/owl-features/.
8. Swarbreck D, Wilks C, *et al*: **The Arabidopsis Information Resource (TAIR): gene structure and function annotation.** *Nucleic Acids Res* 2008, **36:**D1009–D1014.

9. Liang C, Jaiswal P, *et al*: **Gramene: a growing plant comparative genomics resource.** *Nucleic Acids Res* 2008, **36:**D947–D953.
10. McLaren CG, Bruskiewich RM, *et al*: **The International Rice Information System. A platform for meta-analysis of rice crop data.** *Plant Physiol* 2005, **139:**637–642.
11. Lawrence CJ, Harper LC, *et al*: **MaizeGDB: The Maize Model Organism Database for Basic, Translational, and Applied Research.** *Int J Plant Genomics* 2008, **2008:**1–10.
12. Samson D, Legeai F, *et al*: **GénoPlante-Info (GPI): a collection of databases and bioinformatics resources for plant genomics.** *Nucleic Acids Res* 2003, **31:**179–182.
13. Rubin DL, Shah NH, *et al*: **Biomedical ontologies: a functional perspective.** *Brief Bioinform* 2008, **9:**75–90.
14. Chepelev LL, Dumontier M: **Semantic Web integration of Cheminformatics resources with the SADI framework.** *J Cheminform* 2011, **3:**16.
15. Bruskiewich R, Senger M, Davenport G, Ruiz M, Rouard M, *et al*: **The generation challenge programme platform: semantic standards and workbench for crop science.** *Int J Plant Genomics* 2008, **2008:**369601.
16. Goff SA, McKay S, Stapleton AE, Hanlon M, Mock S, Helmke M, Kubach A, Noutsos C, Gendler K, Feng X, Welch SM, O'Meara B, Brutnell T, Leebens-Mack J, Akoglu A: **The iPlant collaborative: cyberinfrastructure for plant biology.** *Front Plant Sci* 2011, **2:**34.
17. Wilkinson MD, Vandervalk B, McCarthy L: *SADI Semantic Web Services – 'cause you can't always GET what you want!* Services Computing Conference, 2009 APSCC 2009 IEEE Asia-Pacific: 7-11 Dec. 2009. 2009:13–18.
18. Gessler D, Schiltz G, *et al*: **SSWAP: A Simple Semantic Web Architecture and Protocol for semantic web services.** *BMC Bioinformatics* 2009, **10:**309.
19. Wilkinson MD, Vandervalk B, McCarthy L: **The Semantic Automated Discovery and Integration (SADI) Web service Design-Pattern, API and Reference Implementation.** *J Biomed Semantics* 2011, **2:**8.
20. Wilkinson MD, Senger M, Kawas E, Bruskiewich R, Gouzy J, Noirot C, Bardou P, Ng A, Haase D, Saiz Ede A, *et al*: **Interoperability with Moby 1.0–it's better than sharing your toothbrush!** *Brief Bioinform* 2008, **9**(3):220–231.
21. Wilkinson M, McCarthy L, *et al*: **SADI, SHARE, and the in silico scientific method.** *BMC Bioinform* 2010, **11:**S7.
22. Spanos D-E, Stavrou P, Mitrou N: **Bringing relational databases into the Semantic Web: A survey.** *Semantic Web* 2012, **3**(2):169–209.
23. *Mapping Relational Data to RDF in Virtuoso.* http://virtuoso.openlinksw.com/dataspace/doc/dav/wiki/Main/VOSSQLRDF.
24. Miles A, Zhao J, Klyne G, White-Cooper H, Shotton D: **OpenFlyData: An exemplar data web integrating gene expression data on the fruit fly Drosophila melanogaster.** *J Biomed Inform* 2010.
25. Cheung KH, Yip KY, Smith A, Deknikker R, Masiar A, Gerstein M: **YeastHub: a semantic web use case for integrating data in the life sciences domain.** *Bioinformatics* 2005, **21**(Suppl 1):i85–96.
26. Lam HYK, Marenco L, Shepherd GM, Miller PL, Cheung K-H: **Using Web Ontology Language to Integrate Heterogeneous Databases in the Neurosciences.** *AMIA Annu Symp Proc* 2006, **2006:**464–468.
27. Bruskiewich R, Davenport G, *et al*: **Generation Challenge Programme (GCP): standards for crop data.** *OMICS* 2006, **10:**215–219.
28. Rahayu JW, Chang E, *et al*: **A methodology for transforming inheritance relationships in an object-oriented conceptual model to relational tables.** *Inf Softw Technol* 2000, **42:**571–592.
29. Tirmizi S, Sequeda J, Miranker D: **Translating SQL Applications to the Semantic Web**. In *Database and Expert Systems Applications. vol. 5181.* Springer Berlin / Heidelberg; 2008:450–464.
30. Dijkstra E: **A note on two problems in connexion with graphs.** *Numerische Mathematik* 1959, **1:**269–271.
31. Pettifer S, Ison J, *et al*: **The EMBRACE web service collection.** *Nucleic Acids Res* 2010, **38:**W683–W688.
32. Kopecký J, Vitvar T, *et al*: **SAWSDL: Semantic Annotations for WSDL and XML Schema.** *IEEE Internet Comput* 2007, **11:**60–67.
33. Bhagat J, Tanoh F, Nzuobontane E, Laurent T, Orlowski J, Roos M, Wolstencroft K, Aleksejevs S, Stevens R, Pettifer S, *et al*: **BioCatalogue: a universal catalogue of web services for the life sciences.** *Nucleic Acids Res* 2010, **38**(Web Server issue):W689–694.
34. Ruiz M, Rouard M, Raboin LM, Lartaud M, Lagoda P, Courtois B: **TropGENE-DB, a multi-tropical crop information system.** *Nucleic Acids Res* 2004, **32:**D364–D367.
35. Flicek P, Amode MR, Barrell D, Beal K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fairley S, Fitzgerald S, Gil L, Gordon L, Hendrix M, Hourlier T, Johnson N, Kahari AK, Keefe D, Keenan S, Kinsella R, Komorowska M, Koscielny G, Kulesha E, Larsson P, Longden I, McLaren W, Muffato M, Overduin B, Pignatelli M, Pritchard B, Riat HS, Ritchie GRS, Ruffier M, Schuster M, Sobral D, Tang YA, Taylor K, Trevanion S, Vandrovcova J, White S, Wilson M, Wilder SP, Aken BL, Birney E, Cunningham F, Dunham I, Durbin R, Fernandez-Suarez XM, Harrow J, Herrero J, Hubbard TJP, Parker A, Proctor G, Spudich G, Vogel J, Yates A, Zadissa A, Searle SMJ: **Ensembl 2012.** *Nucleic Acids Res* 2011, **40:**D84–D90.
36. Goble CA, Bhagat J, Aleksejevs S, Cruickshank D, Michaelides D, Newman D, Borkum M, Bechhofer S, Roos M, Li P, De Roure D: **myExperiment: a repository and social network for the sharing of bioinformatics workflows.** *Nucleic Acids Res* 2010, **38:**W677–W682.
37. Droc G, Périn C, *et al*: **OryGenesDB 2008 update: database interoperability for functional genomics of rice.** *Nucleic Acids Res* 2009, **37:**D992–D995.
38. *Singer.* http://singer.cgiar.org/.
39. Shrestha R, Arnaud E, *et al*: **Multifunctional crop trait ontology for breeders' data: field book, annotation, data discovery and semantic enrichment of the literature.** *AoB Plants* 2010, **2010:**plq008.
40. Eilbeck K, Lewis S, Mungall C, Yandell M, Stein L, Durbin R, Ashburner M: **The Sequence Ontology: a tool for the unification of genome annotations.** *Genome Biology* 2005, **6:**R44.
41. Bizer C, Schultz A: **The Berlin SPARQL Benchmark.** *Int J Semantic Web Inf Syst* 2009, **5:**1–24.
42. *Open Biological and Biomedical Ontologies: current principles.* http://obofoundry.org/crit.shtml.
43. *WebSmatch project: an environment for Web Schema Matching.* http://websmatch.gforge.inria.fr/.

doi:10.1186/1471-2105-14-126
Cite this article as: Wollbrett *et al.*: Clever generation of rich SPARQL queries from annotated relational schema: application to Semantic Web Service creation for biological databases. *BMC Bioinformatics* 2013 **14**:126.