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-Echelles et Extraction de Connaissances en Agronomie : Exemples et Perspectives

Pierre Larmande

IRD
UMR DIADE
Montpellier, France

Version déposée, en attente de validation des rapporteurs

JURY

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

UNIV. MONTPELLIER, ECOLE DOCTORALE I2S, INFORMATIQUE

Résumé

Montpellier, France
UMR DIADE

Habilitation á Diriger des Recherches

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

by Pierre LARMANDE

La compréhension des relations génotype-phénotype est un des axes les plus important de la recherche en agronomie. Or les interactions génotype-phénotype sont complexes à identifier car elles s'expriment à différentes échelles moléculaires dans la plante et subissent de fortes influences de la part des facteurs environnementaux. Les technologies d'analyses haut-débit ne permettent de capturer que partiellement cette dynamique. Même si ces technologies permettent d'aller toujours plus loin dans l'obtention de nouvelles données, notre connaissance reste encore parcellaire pour élucider les mécanismes moléculaires qui régissent l'expression des caractères phénotypiques complexes. Les nouveaux défis consistent à comprendre les relations complexes existant entre les différents éléments moléculaires responsables de l'expression du 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.

Mon projet de recherche 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.

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, l'objectif sera de construire des réseaux d'interaction moléculaires à partir de données éparses afin d'identifier les gènes clés pour l'amélioration des plantes. Plusieurs approches de recherche sont envisagées : intégration des données, enrichissement des connaissances, applications sur les graphes de connaissances.
Dans ce processus, une première voie consistera à transformer et intégrer dynamiquement ces données dans la base de connaissance AgroLD pour les rendre plus facilement utilisables en terme algorithmique. Une deuxième voie consistera à proposer de nouvelles méthodes d'enrichissement des connaissances. Dans un premier temps, en se focalisant sur des méthodes d'annotation sémantique. Puis, afin d'enrichir les liens entre les différents graphes générés et ainsi produire un réseau d'interaction qui permettra la découverte de nouvelles connaissances, de nouvelles méthodes de liage de données seront développées. Enfin, afin de permettre une recherche d'information efficace, plusieurs méthodes et algorithmes de priorisation de gènes candidats seront évaluées et proposées.

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 aider à construire ce projet.

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

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

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

Table des matières

Liste des abréviations

ADN	Acide Désoxyribo Nucléique
ARN	Acide Ribo Nucléique
API	Application Programing Interface
BDR	Base de Donnees 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 Programme
GFVO	Genomic Feature and Variation Ontology
GAF	Gene Ontology Annotation File
GAV	Global As View
GFF	Generic Feature Format
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
RDF	Resource Description Framework
SGBD	Système de Gestion de Base de Données
SNP	Single Nucleotide Polymorphism
SQR	Systèmes de Question-Réponses
URI	Uniform Resource Identifier
VCF	Variant Call Format
XML	eXtensible Markup Language

A Sophie, Nina et Salomé qui m'ont soutenues et encouragées depuis de nombreuses années …

Première partie

CV étendu

Chapitre 1

Curriculum Vitae

1.1 Identité

Pierre LARMANDE
46 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
https://sites.google.com/site/larmandepierre

IRD - UMR DIADE
ICT Lab & LMI RICE USTH
Hanoi, Vietnam
pierre.larmande@ird.fr

Associé au LIRMM
équipe FADO
Montpellier, France
pierre.larmande@lirmm.fr

1.2 Formation

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

D.E.S.S. Informatique Appliquées aux Organisations Université Montpellier 2 :
Obtenu le 2 septembre 2000 à Montpellier, mention assez bien

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

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

Président : Corinne Cauvet, Professeur d'Université Marseille Nord

Rapporteurs : Anne Doucet, Directeur de Recherche CNRS (Lyon) et Christine Froidevaux, Professeur Université Orsay (Paris)

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

Directeurs : Thérèse Libourel, Professeur Université Montpellier 2

1.3 Expérience professionnelle

Février 2001 – Juillet 2002 **Bioinformaticien**, CIRAD, UMR PIA, Montpellier
Financement : projet Génoplante / CDD CIRAD

Aout 2002 – Novembre 2005 **Bioinformaticien, Ingénieur** CIRAD, UMR PIA, Montpellier
Financement : CDD Ingénieur d'Etude CNRS

Décembre 2005 - Septembre 2010 **Bioinformaticien, IE2**, CNRS, Centre d'Ecologie Fonctionnelle Evolutive mis à disposition au CIRAD UMR DAP
Financement : Ingénieur d'Etude permanent CNRS,

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

Septembre 2016 - Octobre 2018 **Bioinformaticien, IE2**, 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.

Novembre 2018 - En cours **Chercheur, CRCN**, 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.

Domaines de recherche : Bio-ontologies, intégration des données et connaissances, Web Sémantique, Génomique, Agronomie

1.4 Investissements au sein de projets scientifiques

— Membre du projet ANR PRCE Data to Knowledge in Agriculture and Biodiversity - D2KAB. 850 K euros. Porteur C. Jonquet.
— Membre du projet international CGIAR – CRP-RICE. 1,5 M euros pour IRD (2017-2022; Co-resp. WP4.5)
— Membre du projet postdoc Labex Numev. Lingua 75 K euros. Porteur C. Jonquet
— Porteur du projet BIOeSAI Spirale IRD 2014-2015 pour le Développement d'une application de gestion de données phénotypique chez le riz. 11 K euros
— Porteur du projet postdoc Labex Numev. LandPan TOGGLE. 2015-2016. 50 K euros (coord. P. Larmande).
— Porteur du projet postdoc Labex Numev. AgroPortal : an ontology repository for agronomy. 2015-2016. 50 K euros (coord. P. Larmande & C. Jonquet).
— Membre du 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. 2012-2017. 2.842 M euros. (Coord. WP5 P. Larmande & P. Valduriez)
— Membre du projet IFB plant node (Institut Francais de Bioinformatique). Développement d'un réseau de ressources bioinformatiques sémantiquement interconnectées. (INRA – CIRAD – CNRS – INRIA - IRD). 2014-2018. 400 K euros. (coord. M. Ruiz)
— Membre du projet ANR Bioadapt Africrop « Documenting African Crop Domestication » Partenaires IRD, CIRAD (Coord Y. Vigouroux) 2013-2017. 698 K euros
— Membre du projet EvoRepRice : « Studying the evolution of reproductive development in the Oryza genus for the improvement of modern cultivated rice ». (coord. S. Jouannic & M. Kater) 2010-2014. 479 K euros
— Membre du projet MENERGEP « Methodologies and new resources for genotyping and phenotyping of African rice species and their pathogens for developing strategic disease resistance breeding programs». Partners : IRD (DIADE), CIRAD (BGPI), Africarice (A. Ghesquière Coord.) projet CGIAR GRISP 800 K$ US

1.5 Responsabilité d'animation de la recherche

Responsabilité d'équipe

Co-Direction ICT Lab USTH - Hanoi

Depuis 2017, j'ai pris la co-direction avec Pr. Luong Chi Mai, du laboratoire mixte IRD-USTH ICT Lab [1] . Il est composé de 11 chercheurs et enseignants chercheurs. Mon rôle comprend en particulier l'animation scientifique, la gestion du budget, les rapports d'activité, la communication. Ces interactions me permettent de développer mon projet de recherche en m'appuyant sur les collaborations au sein du laboratoire.

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 [2]. L'objectif de cet axe est de faciliter l'accès aux données et connaissances en biologie. Il est composé de 10 chercheurs et ingénieurs collaborant sur plusieurs

1. http://ictlab.usth.edu.vn
2. http://www.ibc-montpellier.fr

projets. Mon rôle de coordination comprenait en particulier le suivi des avancements et des livrables, la gestion du budget, les rapports d'activité, la communication. J'y ai également développé de nouvelles méthodes d'intégration sur des données expérimentales. De plus, j'ai supervisé le travail d'un post doctorant, d'un ingénieur et de stagiaires afin de travailler sur les différents livrables.

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

Le plateau *i-Trop* [3] 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. Les missions de ce plateau sont (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 les utilisateurs du plateau. Depuis janvier 2010, j'ai participé au montage et l'animation de cette structure, dont j'ai été le coordinateur en 2012-2013. Je suis actuellement contributeur en termes de services et applications.

Responsable et Membre de Comités d'Organisation

Semantic Web for Biodiversity (S4BIODIV) 2013

S4BIODIV [4] est un workshop attaché à la conférence ESWC2013. J'ai Co-organisé le workshop avec E.Arnaud, C. Jonquet, T. Libourel, I. Mougenot, M. Ruiz. Montpellier, France. Proceedings disponibles sur CEUR [5]

PhenoHarmonis : Harmonization, semantic and interoperability of phenotypic and agronomic data Workshop 2014 - 2016 - 2018

Suite au succés du workshop S4BIODIV, le groupe d'organisation a travaillé sur cette nouvelle série. J'ai co-organisé PhenoHarmonIS [6] avec E. Arnaud, M. Ruiz, P. Neveu, C. Pommier, D. Pot, JF Rami. Montpellier, France.

IC2016 : 27e Journées francophones d'Ingénierie des Connaissances 2016

6-10 juin. Montpellier, France. J'ai pu co-organiser la conférence en recherchant des financements permettant d'inviter des keynotes speakers. [7]

AgroHackathon : discovering AgroPortal & AgroLD. 2016

Premier Hackathon [8] dédié à l'intégration de données agronomiques. Co-organisation avec C. Jonquet. Montpellier, France.

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 des données dans la communauté scientifique. Je coordonne depuis janvier 2017, le groupe pour le riz. l'objectif Rice Data Interoperability WG [9] sera de

3. http://bioinfo.mpl.ird.fr
4. http://semantic-biodiversity.mpl.ird.fr
5. http://ceur-ws.org/Vol-979/
6. https://tinyurl.com/PhenoharmonIS2018
7. https://ic2016.sciencesconf.org
8. https://www.meetup.com/AgroHackathon
9. https://www.rd-alliance.org/groups/rice-data-interoperability-wg.html

proposer l'utilisation de standards et un guide de bonnes pratiques pour échanger et publier les données produites sur le riz.

1.6 Activités d'enseignement

— Enseignement de Web Sémantique au Master 1 ICT de l'Institut Francophone d'Informatique parcours 1 et 2, Hanoi, 2018-2019 (60h - 2x30h)
— Enseignement de Bioinformatique au Master 2 ICT USTH, & Bio, 2017-2018 (50h)
— Enseignement au Master 2 ICT USTH, Systèmes d'information Geographique, 2017 (25h)
— Enseignement Master BioPharma USTH (Hanoi), Module Bioinformatique 2013 (40h)
— Enseignement IUT Informatique, UMII, TP Base de données, 2004-2006 (120h au total)
— Enseignement au DESS de Bioinformatique, UMII, TP BioPerl, 2002-2003 (50h)

1.7 Encadrements

Thèses

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 : Dec. 2011
— Situation actuelle : Post-Doctorant au Swiss Institute of Bioinformatics. Auparavant Post-doctorant au CNRS Roscoff.
— Financement : Bourse Région Languedoc Roussillon - CIRAD

Stages Post-doctoraux

J'ai collaboré avec 3 docteurs en stages post-doctoraux et 2 ingénieurs de recherche.

+ [2015 – 2017] N. El Hassouni – Contribution dans le développement du projet AgroLD.
  Financement INRA sur le projet IFB
+ [2015 - 2017] A. Toulet – Contribution au développement d'un portail d'ontologies pour l'agronomie basé : AgroPortal.
  Financement Numev puis IBC
  o Co-supervision avec Clément Jonquet
+ [2014 –2016] G. Sempéré – Conception et développement de l'application Gigwa.
  Financement Cirad.
+ [2014 - 2016] : A. Venkatesan – Intégation de données utilisant les métadonnees et ontologies pour pour agréger les données de plusieurs ressources hétérogenes.
  Financement IBC
+ [2012-2013] J.Wollbrett - Automatiser l'intégration de bases de données relationnelles distribuées à travers l'enrichissement sémantique de vues RDF avec BioSemantic –
  Financement Cirad - IBC

Masters

J'ai encadré 20 masters.

Encadrement de stages de master professionnel

- 2001 C. Tranchant – Développement d'un système d'information sur la traçabilité des échantillons OGM . DESS IAO UMII
  Situation suivante : Ingénieur Bioinformatique, IRD
- 2002 G. Droc – Développement d'un pipeline de traitement des séquences génomiques pour les mutants d'insertion chez le riz – Master2 Bioinformatique UMII
  situation suivante : Ingénieur Bioinformatique, Cirad
- 2014 F. Philippe - Analyse de données de variations génétique dans les riz – Master2 bioinformatique Lumini
  co-encadrement : G. Sempere – Cirad
  Situation suivante : Ingénieur Bioinformatique, INRA
- 2016 A. Petel – Contribution au développement de Gigwa – Master2 Polytech Grenoble
  co-encadrement : G. Sempéré
  Situation suivante : Volontaire International Cirad, la Reunion
- 2016 D. Hyżorek - Epigenetic Data Integration and Analysis – Master2 parcours BCD UM
  co-encadrement : M. Mirouze
  Situation suivante : Chercheur en Pologne
- 2017 B. Vautrin – Développement de module ETL pour l'intégration de ressources dans AgroLD. PolyTech. (stage international - Hanoi)
  Situation suivante : Dernière année polytech
- 2017 J-C Idjellidaine – Propososition d'un systeme de recommandation pour valider les mappings sémantiques dans AgroLD. L3 informatique
  co-encadrement : N. El Hassouni
  Situation suivante : M1 AIGLE

Encadrement de stages de master recherche

- 2007 S. Fromentin - Développement d'un Framework de services web pour l'interopérabilité de ressources agronomiques - Master2 Bioinformatique Orsay
  Situation suivante : Consultant Bioinformatique SS2I
- 2008 J. Wollbrett - Intégration automatique d'une ontologie de domaine dans un annuaire de service web bioinformatique : Biomoby – Master2 Bioinformatique UMII
  co-encadrement : M. Ruiz – Cirad
  Situation suivante : Thèse CIRAD-Région LR
- 2014 G. Tagny (M1) - Développement d'une base de connaissances sur les gènes régulateurs de la ramification chez le riz – Master1 Institut Francophone d'Informatique – Hanoi
  Situation suivante : Master 2
- 2014 L. Le Ngoc (M1) - Développement d'une application de gestion de données phénotypique chez le riz – Master1 Institut Francophone d'Informatique – Hanoi
  Situation suivante : Master 2
- 2015 I. Chentli - Facilitation de l'accès aux données biologiques sémantiquement structurées – Master2 parcours BCD UM
  co-encadrement : K. Todorov
  Situation suivante : Ingénieur Bioinformatique, IMGT-CNRS
- 2015 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 – Hanoi
  Co-encadrement : Pascal Neveu – INRA
  Situation suivante : Thèse CIFRE Crédit Agricole, Brest

— 2015 G. Tagny (M2) - The Agronomic Linked Data (AgroLD) project. Master2 Institut Francophone d'Informatique – Hanoi
Co-encadrement : A. Venkatesan
Situation suivante : Thèse Ecole des Mines Ales, Nîmes
— 2016 S. Remini - Acquisition automatique de connaissances à partir de textes scientifiques – Master2 parcours BCD UM
co-encadrement : K. Todorov
Situation suivante : Recherche d'emploi
— 2016 S. Zevio - Indexation de données issues du web sémantique dans le domaine agronomique – Master1 parcours DECOL UM
Situation suivante : Master 2 DECOL UM
— 2017 A. Diouf – Proposition et implémentation d'algorithmes de liage de données RDF dans AgroLD – M1 BCD
co-encadrement : K. Todorov
Situation suivante : M2 BCD
— 2018 A. Sayadi – Liage de données complementaires dans le contexte d'AgroLD – M2 AIGLE
co-encadrement : K. Todorov
Situation suivante : Recherche d'emploi
— 2018 H. Do – Evaluating Name-Entity Recognition approaches in plant molecular biology – M1 USTH
co-encadrement : K. Tanh
Situation suivante : M2 USTH
— 2018 K.M. Djibril – Developping an Ontology Matching workflow using AgroPortal API – M2 USTH
Situation suivante : Job IT

1.8 Autres implications

Relectures d'articles

— Nucleic Acids Research,
— Databases,
— Bioinformatics,
— BMC Bioinformatics,
— Current Plant Biology

Comité de programme

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

JURYS

— Rapporteur de stages de M2 Bio-Informatique (UM) (régulièrement depuis 2002)
— Jury Masters BioPharma USTH 2017
— Jury de concours CNRS (Ingénieur d'Etude)

Expertises

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

1.9 Prototypage

— AgroLD [10] – 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 [11] – Développement d'une base de données génomiques, constructeur de requêtes, API de services web. Co-encadrement G. Sempere.
— BIOeSAI [12] - Développement d'une base de données phénotypique, constructeur de requêtes
— BioSemantic [13] - Développement d'un constructeur de requêtes SPARQL au-dessus de bases de données relationnelles biologiques. Co-encadrement M. Ruiz
— OryGenesDB [14] - Développement d'une base de données génomique pour le riz.
— Oryza Tag line [15] - 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)

10. http://www.agrold.org
11. http://southgreen.fr/content/gigwa
12. http://vmbioesai-dev.ird.fr:8080/Syspherice
13. http://www.southgreen.fr/content/biosemantic-tool
14. http://orygenesdb.cirad.fr
15. http://oryzatagline.cirad.fr

Chapitre 2

Liste des publications

Je publie dans le domaine de l'Intégration de données et de connaissances et dans le domaine de la bioinformatique. La plupart des publications sont indexées par :

— DBLP Computer Science Bibliography (17 entrées le 20/02/2019) :
http://dblp.uni-trier.de/pers/hd/l/Larmande:Pierre

— PUBMED (14 entrées le 20/02/2019) :
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 20/03/2019. Le rang des conférences est issu du site CORE (http ://103.1.187.206/core) et mis à jour le 20/03/2019. 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 proceedings. 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.

Thèse

Sujet : Mutualiser et partager, un défi pour la génomique végétale
Date de soutenance : Le 20 Décembre à Montpellier, mention très honorable
Université : Montpellier 2, école doctorale informatique
Président : Corinne Cauvet, Professeur d'Université Marseille Nord
Rapporteurs : Anne Doucet, Directeur de Recherche CNRS
Christine Froidevaux, Professeur d'Université Orsay
Encadrants : Isabelle Mougenot, Maître de Conférence Universite Montpellier 2
Manuel Ruiz, Chercheur Cirad (Montpellier)
Directeur : Therese Libourel, Professeur d'Université Montpellier 2

Publications nationales avec comité de lecture

Édition d'ouvrages

E1 Do H, Than K, **Larmande P**.
Evaluating Named-Entity Recognition approaches in plant molecular biology. MIWAI Vietnam (Hanoi). Springer LNAI proceedings 11248. pp 219-225 2018

E2 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

Publications internationales avec comité de lecture

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. Yaw Nti-Addae, Dave Matthews, Victor Jun Ulat, Raza Syed, Guilhem Sempere, Adrien Petel, Jon Renner, **Pierre Larmande**, Valentin Guignon, Elizabeth Jones, Kelly Robbins.
Benchmarking Database Systems for Genomic Selection Implementation. 2019. Gigascience (Accepted). Impact Factor : 7.31
3. 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. Bio-Informatics. pii :btz190. Impact Factor : 5.41
4. 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
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. 2018 (In Press) Impact Factor : 7.31
6. 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
7. 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

8. 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 : 5.134
9. 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
10. 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.
11. 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
12. 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
13. 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
14. 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
15. 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
16. 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
17. 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
18. **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

19. 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
20. 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
21. 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
22. 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 avec comité de lecture

C1 **Larmande P**.
The AgroLD project A Knowledge Graph-based Semantic Database for rice functional genomics. Oral presentation at International Symposium on Rice Functional Genomics ISRFG 2018. Tokyo (Japan) 2 pages.

C1b Do H., Than K., and **Larmande P**.
Evaluating Named-Entity Recognition approaches in plant molecular biology. MIWAI 2018.Proceedings LNCS AI ; 2018. 14 Pages

C1c Do H., Than K., and **Larmande P**.
Comparative NER approaches in plant molecular biology. CiCling 2018. Proceedings RCS ; 2018 (In Press) 7 Pages

C1d **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). 2 pages

C2 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

C3 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

C4 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) 2 pages

C5 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.

C6 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. ACM. p. 27 :1- :9.. Nimes (France)

C7 **Larmande P**.
Ontology-based services and knowledge management in the Agronomic Domain. Oral presentation at the 6th Research Data Alliance Conference RDA'2015. Paris (France). 2 pages.

C8 **Pierre Larmande**.
Gigwa - Genotype Investigator for Genome Wide Analyses. Computer demo at Plant and Animal Genomes Conference PAG 2015. San Diego, (USA). 2 pages.

C9 **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.

C10 **Larmande P.**, Mougenot I., Jonquet C., Libourel T., Ruiz M., Arnaud E.
Proceedings Semantics for Biodiversity Workshop. ESWC 2013. Montpellier (France) 4 pages.

C11 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.

C12 Julien Wollbrett, **Pierre Larmande** and Manuel Ruiz.
Towards Automatic Generation of Semantic Web services for relational Databases. Oral presenation at the International Workshop on Resources Discovery in conjunction with ESWC 2011. Heraclion (Greece) 6 pages.

C13 **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.

C14 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.

C15 Fromentin S., Droc G. and **Larmande P**.
A personalized, integrated system for rice functional genomics. Poster at Network Tools and Applications in Biology NETTAB 2007, Pise (Italy) 4 pages.

Communications nationales avec comité de lecture

C16 **Larmande P**.
Gigwa : Genotype Investigator for Genome Wide Analyses. JOBIM 2018. Marseille. 2 pages

C16a **Larmande P**.
Exposing French agronomic resources as Linked Open Data. Oral Presentation Conference Francophone d'ingénierie des connaissances, IC 2016. Montpellier (France) 6 pages.

C17 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.

C18 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. Presentation orale ā la Conference Francophone d'ingenierie des connaissances, Rennes, 2015. 6 pages.

C19 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.

C20 Le Ngoc L., Jouannic S. and L**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.

C21 Wollbrett J., **Larmande P**. and Ruiz M.
Intégration automatique d'une ontologie de domaine dans un annuaire Biomoby. Presentation orale aux Journées ouvertes pour la Biologie, l'informatique et les Mathematiques JOBIM2009, Nantes (France). 8 pages.

C22 **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. 8 pages.

Deuxième partie

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

Chapitre 3

Contexte scientifique et problématique

Dans ce chapitre nous allons rappeler tout d'abord les enjeux actuels de la biologie moléculaire chez le riz puis les problématiques des domaines concernés au confluent de la biologie et de l'informatique ainsi qu'un état de l'art autour des principales avancées existantes. [1]

3.1 Les enjeux actuels de la biologie moléculaire chez le riz

3.1.1 Le séquençage des génomes de riz

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 repartie sur les chromosomes) permettrait de comprendre de l'ensemble des mécanismes biologiques.

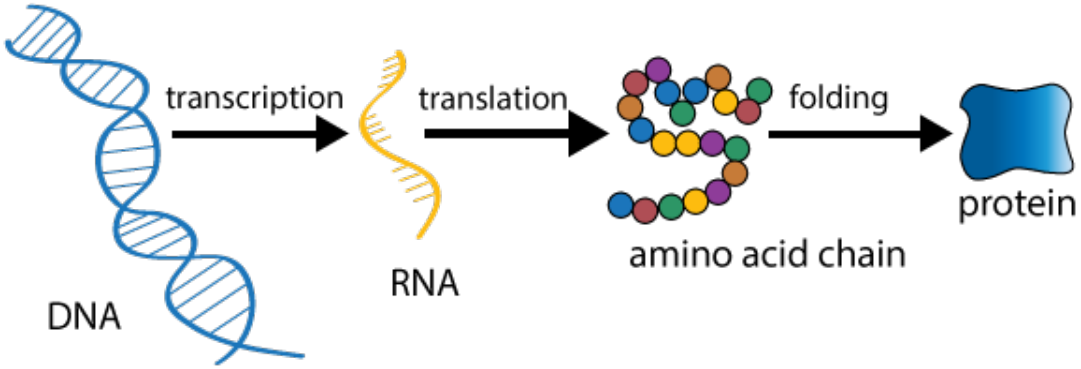


FIGURE 3.1 – Le dogme central de la biologie moléculaire indique que l'information contenue dans l'ADN génomique est successivement transformée en ARN, chaîne amino acides et protéine. Crédits :biosocialmethods.isr.umich.edu

Cette hypothèse, 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 projet de séquençage pour les plantes cultivée. Oryza sativa est un génome diploide de type AA qui comprend deux sous-espèces principales (voir figure 3.2) : la variété japonica à grain court et collant, et la variété de riz indica à grain long et non collante. 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

1. Les questions ont été parfois déjà abordées dans le chapitre 2, mais nous souhaitions les rappeler sous un angle plus général.

sa longueur totale de 389 Mega-bases fut achevée en 2004. La séquence génomique de haute qualité servit pendant de nombreuses années de modèle aux projets de séquençage d'autres cultures céréalières possédant de grands génomes et des contenus chromosomique complexes [117]. 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. Indica fut séquencé 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 réalisé par le TIGR (The institute of Genome Research) et aujourd'hui gérée par la Michigan State University (MSU) [2]. 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 [3]. 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 de leur données.

Dans les années qui ont suivies, 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 disputer les gènes par ciblages spécifiques (cf.Transgénèse [4]) telles que celle décrites dans la section 4.2 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 année 2000 avec pour objectif de séquencer et étudier la structure évolutive des génomes diploïdes du groupe AA et BB [199].

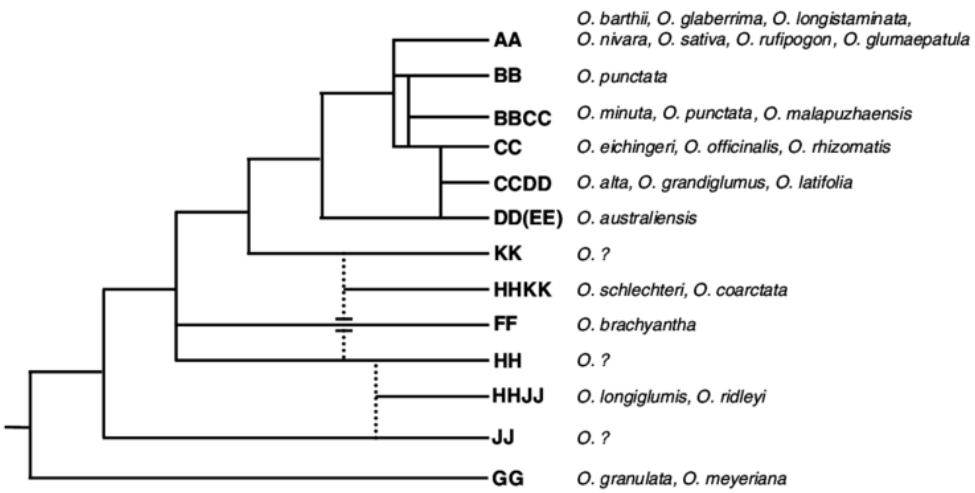


FIGURE 3.2 – Arbre Phylogénétique du genre Oryza (Modifié à partir de Ge et al. 1999). Crédits :Projet OMAP

2. http://rice.plantbiology.msu.edu/cgi-bin/gbrowse/rice/
3. http://rapdb.dna.affrc.go.jp
4. https://fr.wikipedia.org/wiki/Transg%C3%A9n%C3%A8se

3.1.2 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 [191]. Ce projet, a utilisé une collection de base de 3 000 accessions de ressources génétiques 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 ressource fondamentale 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.

Les études GWAS ou 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 donnée (trait). Les marqueurs polymorphes les plus couramment utilisés pour 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 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 et l'effet de chaque QTL. Comme l'illustre la figure 3.3, 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.

3.1.3 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 ont identifié de nombreux variants génétiques associés à des maladies complexes ou à d'autres caractères phénotypiques. Toutefois, même si ces découvertes enrichissent grandement nos connaissances sur les bases génétiques de la variation phénotypique, la plupart des variantes identifiées jusqu'à présent n'expliquent qu'une faible proportion des facteurs génétiques causaux, laissant à découvrir et expliquer l'héritabilité restante [115]. Par ailleurs, même avec une compréhension complète

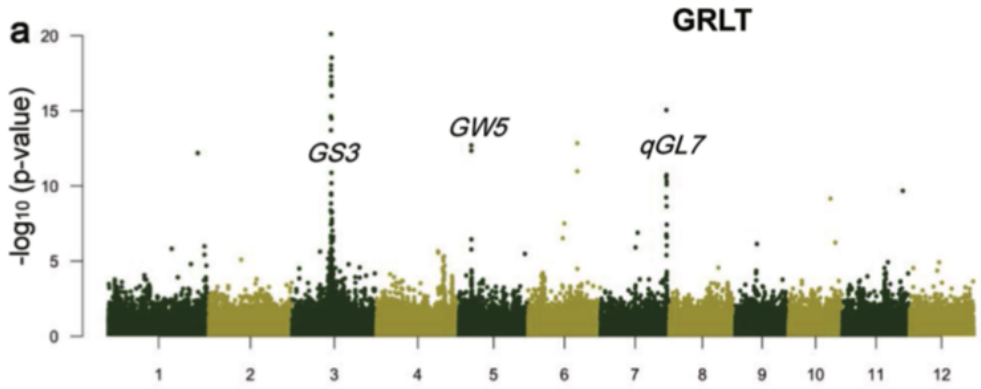


FIGURE 3.3 – Analyse GWAS réalisée pour la longueur du grain (GRLT) chez Oryza sativa (Modifié à partir de Wang et al. 2018). Illustration d'un manathan plot montrant la corrélation entre des variants et le caractère de longueur du grain. Ici chaque point représente un SNP avec sur l'axe des abscisses 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

de la génétique d'un trait 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 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-GWAS» [58, 78].

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 carte génotype – phénotype ne peut être établie sans données phénotypiques détaillées [79]. 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. 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.

3.1.4 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 3.4, la régulation de l'expression des gènes conduisant à un phénotype peut intervenir à

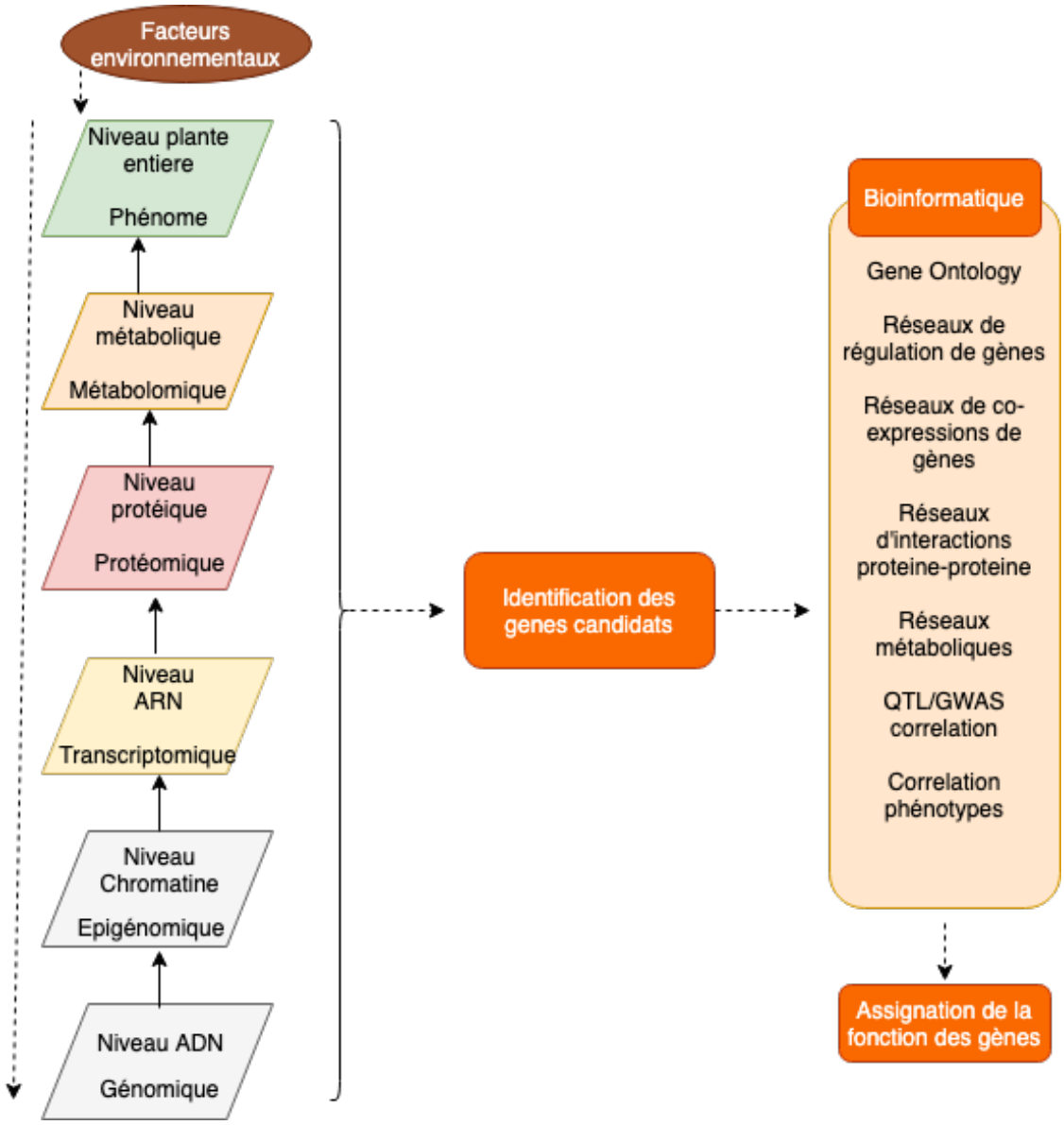


FIGURE 3.4 – Différentes échelles de la régulation de l'expression des gènes conduisant à un phénotype [28]

différents niveaux au sein d'une cellule et de l'organisme.

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

— Au niveau de **l'épigénome** - ensemble des propriétés physico- chimique 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 leur lien 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 3.5[5], des modifications chimiques de la chromatine permettent de libérer l'accès à l'ADN et favorisent l'expression des gènes.

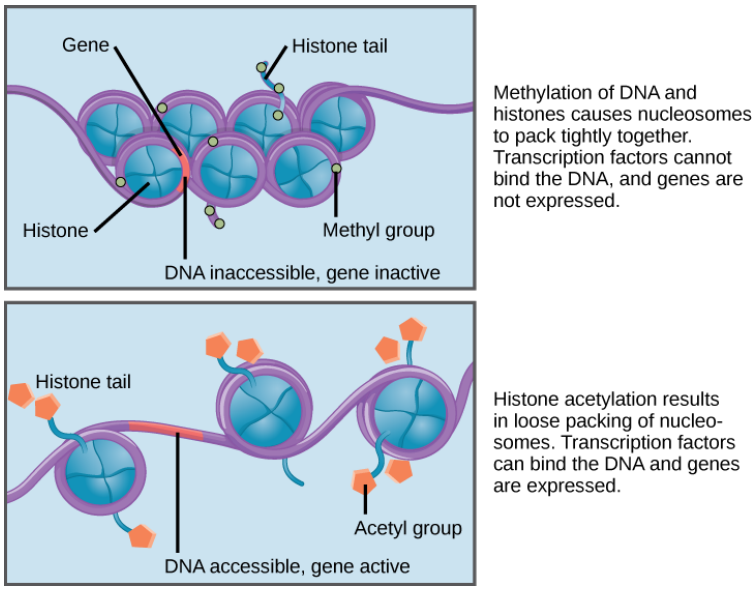


FIGURE 3.5 – Mécanisme d'ouverture du nucléosome - complexe histones-ADN - pour permettre la transcription des gènes. Crédits :Lumen

— **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 toutes les espèces 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 a 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 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 acquises sur ces réseaux constituent une base importante pour la compréhension de la biologies des systèmes.

5. https://courses.lumenlearning.com

— **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.

3.2 L'intégration de données en biologie

3.2.1 L'hétérogénéité des systèmes

Une meilleure compréhension des relations génotype-phénotype nécessite une intégration de données biologiques de diverse nature. Or, une caractéristique de la biologie moléculaire est que le volume et la variété des données produites par les technologies haut-débit ont une croissance exponentielle. 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 [148]. Chaque année, de nouvelles bases de données moléculaires 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. Toutefois, ne dérogeant pas aux règles de financement des projets de recherches académiques qui sont en moyenne de 3 à 5 ans, on constate que la plupart de ces systèmes ont une durée de vie courte. En effet, même s'ils sont toujours disponibles sur Internet des années plus tard, leurs mises à jour ou évolution sont souvent stoppées. La seule chance de survie d'un nouveau système est de trouver de nouveaux financements, d'avoir un modèle économique incluant des services payants ou de créer une entreprise. Deux importantes ressources Tair et Uniprot dans le domaine ont expérimenté plusieurs modèles de financement pour financer leur fonctionnement et en font un bilan [146, 59]. Mise à part quelques exemples comme ceux-ci, la situation reste compliquée dans la plus part des cas. À ce stade, il est important de noter que la qualité des données présentées par ces systèmes via Internet doit être garantie par chaque fournisseur de données ou du système d'information. Jusqu'à présent, aucune norme de qualité n'a été définie pour la mise en œuvre de ces données. La provenance des données est un élément important à prendre en compte dans les analyses réalisée à partir de données issues de plusieurs ressources distribuées [32].

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 énumèrent les formes d'hétérogénéité rencontrées que nous avons adopté [106] :

— **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 "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.

Également, l'une des classifications les plus complètes provient de Pluempitiwiriyawej et Hammer, "Classification Scheme for Semantic and Schematic Heterogeneities in XML Data Sources" [15]

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. 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 moléculaires distribuées et hétérogènes constitue une partie importante de la recherche en bioinformatique.

3.2.2 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 [188].

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.

La première approche 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. *L'intégration matérialisée* 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

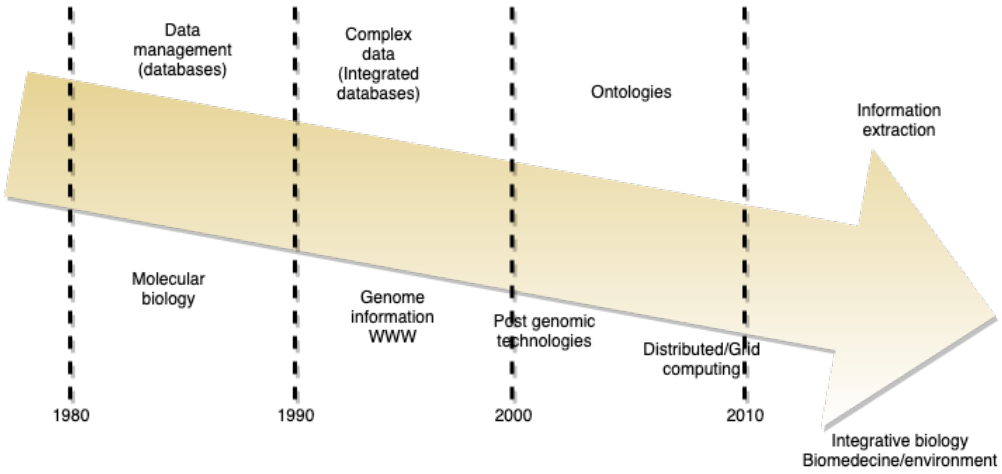


FIGURE 3.6 – Évolution des systèmes d'information en parallèle des méthodes biologiques. Crédits : Valencia 2002 [188]

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 [193, 194]. Généralement, les données sont situées sur différents systèmes réparties 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 parles 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) [70]. 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 [104].

Une synthèse des principales approches d'intégration en bioinformatique réalisées au cours des dernières années a été discutée dans Cohen-Boulakia et Leser [33]. 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 [52], Entrez [137]).

— **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 [43], DiscoveryLink [68]).

— **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 [43], Atlas [163], BioWarehouse [87], Columba [185]).

— **Les boîtes à outils** qui facilitent la construction de tels entrepôts de données sont très populaires. Parmi elles, BioDWH [186] GMOD-CHADO [204], BioMART [168], Intermine [170], Tripal [56, 154, 35].

— **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) [172] et ONDEX [95, 174, 175]).

— **Systèmes hybrides** s'appuyant sur plusieurs technologies. Par exemple, Biozon [17] 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 [61, 67, 124].

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

Récemment les développements de nouvelles générations de base de données NoSQL, ont ouverts de nouvelles perspectives notamment en bioinformatique. Dans un premier temps, des applications dans le domaine génomique ont été développées avec CouchDB [116, 6], Cassandra [60] et MongoDB [159]. 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 [135]. Par ailleurs, le développement de framework d'analyses de données volumineuses en parallèle tels que Hadoop ou Spark, a donné lieu à des applications [156, 133, 176].

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 [74, 112]. Ainsi, nous constatons le développement d'un nombre croissant de base de données de graphes [73, 140]. 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 [197]) et l'interopérabilité des données [49, 105]. Elixir-Europe est une structure européenne impliquant les principaux instituts publics nationaux qui favorise le développement de services en Bioinfomatique et favorise la dissémination 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 [16] 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 [14], OpenPHACTS [198] et EBI RDF [85]. Toutefois, il n'y a pas d'équivalent dans le domaine agronomique.

3.3 Représentation des données

3.3.1 Rappel sur le web de données

Les origines du web

La recherche autour de l'organisation et l'accès à l'information sont des thématiques qui ont émergé bien avant les avancées technologiques que nous connaissons aurjoud'hui. Paul Otlet [139] (Otlet, 1934) fut l'un des premiers à conceptualiser la science de l'information et à imaginer une machine le *le Mundaneum* qui ressemble à l'Internet aujourd'hui. Vannevar Bush [189] proposa dix ans plus tard un modèle d'accès aux connaissances en inventant une machine imaginaire *le Memex* permettant de gérer toute l'information produite par un individu. En 1965, Ted Nelson [128] (Nelson, 1965) proposa sa vision pour la gestion de documents en définissant le concept de lien hypertexte qui sera concrétisé 25 ans plus tard lors de la création du premier serveur Web. La même année, Margaret Dayhoff proposa le premier atlas de séquences et structures protéiques, prémisse des premières grandes bases de données de biologie moléculaire et de la bioinformatique. Le développement de nouvelles technologies dans différentes disciplines telles que les mathématiques, la physique, la biologie et la médecine auront permis de grandes avancées dans la représentation

des connaissances. N'est ce pas grâce à ses travaux au CERN que Tim-Berners-Lee[7] proposa la mise en place du Web en 1989 ?

Dans les années 1990, la mise en place du Web a permis de mettre en place une stratégie de partage de ressources sur un réseau de machines. Cette stratégie s'appuie sur 4 principaux dispositifs technologiques.

— un langage d'encodage des documents basé sur le Standard Generalized Markup Language (SGML) ; un langage à balises proche de HTML,
— un protocole de communication HyperText Transfer Protocol (HTTP) pour lier une machine client à un serveur,
— un mécanisme d'identification Uniform Ressource Identifier (URI) pour référencer de façon unique n'importe quelle ressource sur le web,
— une relation entre les documents sous forme d'hypertextes pour lier différentes données.

Les langages du Web de données

Plus tard, le World Wide Web Consortium (W3C) est créé par Tim Berners-Lee où différentes recommandations sont proposées pour normaliser et rendre compatible les différentes technologies du Web et les services pour tous. Les différents standards proposés sont l'HyperText Markup Language (HTML) pour l'écriture et la mise en page Internet ainsi que HTTP comme protocole d'échanges entre les machines des utilisateurs et les serveurs. Cependant, le partage de données nécessite différents dispositifs technologiques. A cette époque déjà, les chercheurs imaginaient un Web à travers lequel les machines seraient capables d'analyser le contenu des données et d'interagir avec l'Homme.

A la même période, soutenu par le W3C, le Web Sémantique se met en place pour composer le nuage d'informations que nous connaissons. 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. Alors qu'un URI n'autorise uniquement que les caractères ASCII, son extension, un IRI, autorise les caractères internationaux ainsi que l'identification d'une entité dans plusieurs langages.

Parmi les technologies utilisées pour exposer des données sur le Web de données RDF, RDFS, OWL et SPARQL sont les éléments importants.

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. RDF peut être représenté selon différentes syntaxes. L'une des syntaxes (ou sérialisation) de ce langage est RDF/XML. D'autres syntaxes sont apparues ensuite, cherchant à rendre la lecture plus compréhensible. Citons par exemple N3, Turtle, RDFa et plus récemment JSON-LD.

Le RDF et le **RDFS (Resource Description Framework Schema)** sont considérés comme les premières fondations de l'interopérabilité sémantique. A la différence du RDF, le RDFS fournit

7. Retrouvez plus en détail l'histoire du web dans le livre sur le Web Semantique [62], de F. Gandon, C. Zucker, O. Corby, 2012, Dunod

des éléments de base pour la définition d'ontologies ou vocabulaires 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.

Le language OWL (Web Ontology Language) étend le langage RDFS en offrant une meilleure expressivité pour définir des ontologies. Il utilise en effet, une richesse plus importante dans la manière de décrire les concepts et leurs relations à travers son vocabulaire et sa manière de décrire les contraintes en utilisant des propriétés de description de classes telles que des cardinalités, unions, intersections et complémentarités. Lors de la conception d'un ontologie, il est possible d'étendre ou d'intégrer différentes ontologies existantes. Il sera alors nécessaire d'indiquer de quels vocabulaires les termes employés proviennent. Une ontologie est composée de classes, de propriétés et d'instances. Les classes définissent un groupe de sujets ayant des caractéristiques similaires hiérarchisés avec une possibilité d'héritage multiple. Les propriétés expriment des faits sur des individus. Par exemple la position de début et de fin d'un gène et sa localisation chromosomique. Les instances quant à elles, énoncent un axiome d'appartenance à une classe ou concernant l'identité des sujets. Nous ne détaillerons pas plus dans ce mémoire. Grigoris Antoniou et Frank Van Harmelen apportent de plus amples informations sur le langage OWL [8].

Le langage de requête SPARQL 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 5.3.2 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 sont écrites 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).

3.3.2 Exemples de représentation des données en biologie

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

Comme illustré à la figure 3.7, les graphes sont constitués de sommets et d'arêtes représentant les ressources et les prédicats respectivement. 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 d'un littéral ou d'une ressource respectivement.

La figure 3.7 représente : le sujet par l'URI ou la ressource
http ://identifiers.org/ensembl.plant/Os01g0100100, le prédicat indique que ce sujet possède une position chromosomique, — la valeur de l'objet est **chromosome 1**.
La figure 3.8 représente le triplet donné en exemple 3.7 dans le contexte du schéma RDF développé pour AgroLD (nommé agrold_vocabulary). La ressource identifiée par **ensembl :Os01g0100100** est décrite comme appartenant à la classe **Gene** du vocabulaire AgroLD par la relation **rdf :type** qui 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**.

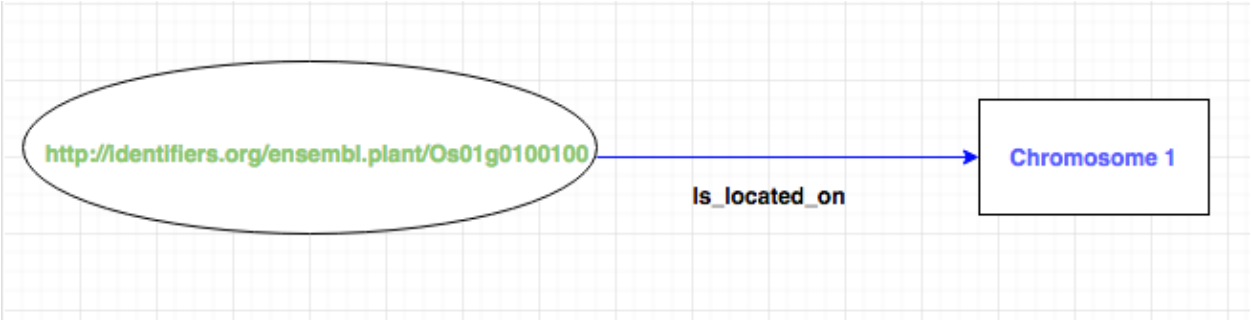


FIGURE 3.7 – Représentation d'un triplet RDF (sujet, prédicat, objet)

Nous avons vu qu'il était possible d'insérer des informations en triplets dans un graphe hiérarchisé pour lier, représenter et publier les données sur le Web. Les ressources peuvent être décrites selon un vocabulaire bien déterminé tels que les noms de classes, les types de ressources et les types de relations entre elles. Cependant, le 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.

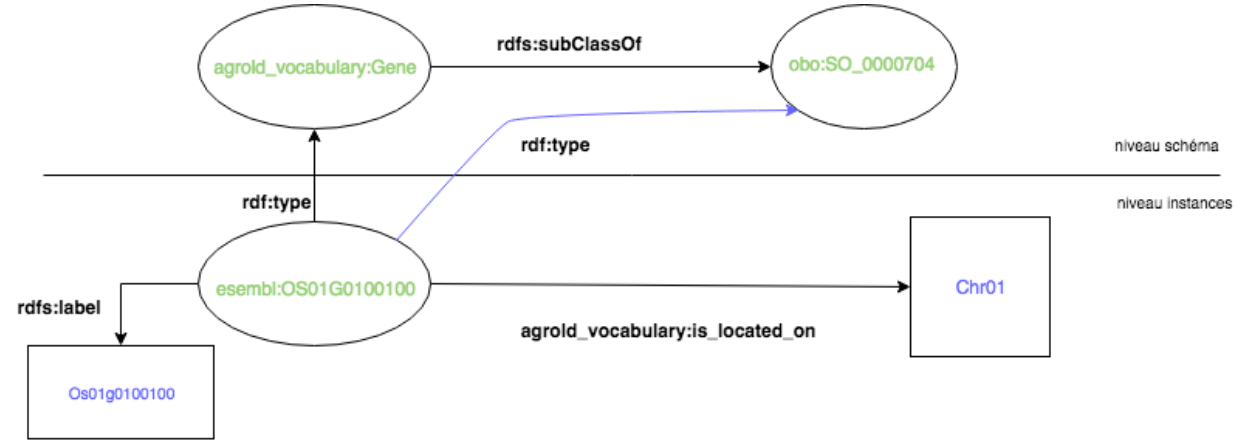
De grands efforts ont été déployés depuis plus de 10 ans afin de structurer et partager les vocabulaires au sein de la communauté Biomédicale et des sciences de la vie. Parmi les initiatives les plus importantes citons MGED (Micro Array Gene Expression Data) [192] qui décrit les données d'expériences de puces à ADN mais surtout OBO (Open Biomedical Ontologies) [169, 66, 180] 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 Ontology's (NCBO) développe et maintient Bioportal [134] une plateforme qui contient plus de 400 ontologies et terminologies biomédicales. Ontobee [136] permet de faciliter le partage d'ontologies, l'intégration et l'analyse des données ainsi que leur visualisation et leur requêtes. Le Ontology Lookup Service (OLS) [38] 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'il vise à faire participer à la fois les utilisateurs et les machines [16].

Dans le Web sémantique, il existe différentes bases de connaissances qui sont regroupées en

fonction du domaine d'expertise. MeSH (Medical Subject Headings) [8] est une terminologie biomédicale très utilisée dans le domaine scientifique et également transformée pour le Web de données. MeSH est le trésor de termes de référence permettant d'indexer, rechercher et classer des documents tels que ceux de PubMed et bien d'autres bases de données dans le domaine biomédical. Aujourd'hui, c'est la ressource la plus utilisée et reliée aux autres ressources. En 2009, il a été recensé pour le domaine des sciences de la vie 191 millions de "liens sortant" 4 et plus de 3 milliards de triplets RDF soit environ 10% du total. Aujourd'hui, il en représente 30% pour tous les domaines confondus. De plus, l'ensemble de jeux de données liées a été multiplié par 8 en l'espace de 10 ans : il passe de 41 à 332 sur un total de 1163. Ces données liées structurées sont consultables directement en ligne et téléchargeables [9]. Elles sont pour la majorité mises à jour régulièrement. Max Schmachtenberg et ses collaborateurs ont recensé en avril 2014 l'état des ressources disponibles et plus récemment Andrejs Abele et al. Une vue d'ensemble est présentée en fonction des domaines, du nombre d'ensembles de données, de triplets et des liens les reliant [155].

8. https://www.nlm.nih.gov/mesh
9. http://lod-cloud.net




```
@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

3.4 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. 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.

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** [10], **BioNLP** [11] 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 [13] : 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 bioinformatique 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 [76, 63].

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 [57]). 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 [25, 30]. 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 2 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 [143] et les CRF (Conditional Random Fields) [96]. Les CRF sont devenus le modèle standard de facto [162], étant la méthode de choix pour la quasi-totalité des outils ayant remporté des compétitions récentes de type NER, comme BioCreative IV [92] ou i2b2 [187] . Les outils NER populaires utilisant les CRF sont, par exemple, ABNER (A Biomedical Named Entity Recognizer) [161] et BANNER [102].

10. Biocreative - http://www.biocreative.org
11. http://2016.bionlp-st.org

Les méthodes hybrides combinent des méthodes d'apprentissage automatique avec des techniques basées sur des dictionnaires ou des règles. Par exemple, ChemSpot [150] intègre les résultats d'un modèle CRF avec un module d'appariement de dictionnaire pour NER chimique.

Récemment, l'utilisation d'approches de réseaux de neurones combinés aux CRF dans l'analyse de texte montrent des résultats bien meilleurs qu'avec les approches précédentes [158]. Notamment, les modèles LSTM-CRF (Long Short Term Memory model combinés avec Conditional Random Fields) [98] offrent des résultats encourageants. Cependant, ces méthodes nécessitent un volume de données important afin d'optimiser les phases d'entraînements [69].

Chapitre 4

Synthèse des activités de recherche et résultats obtenus

4.1 Préambule et 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 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 bio-molécule. 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 l'organisation 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.

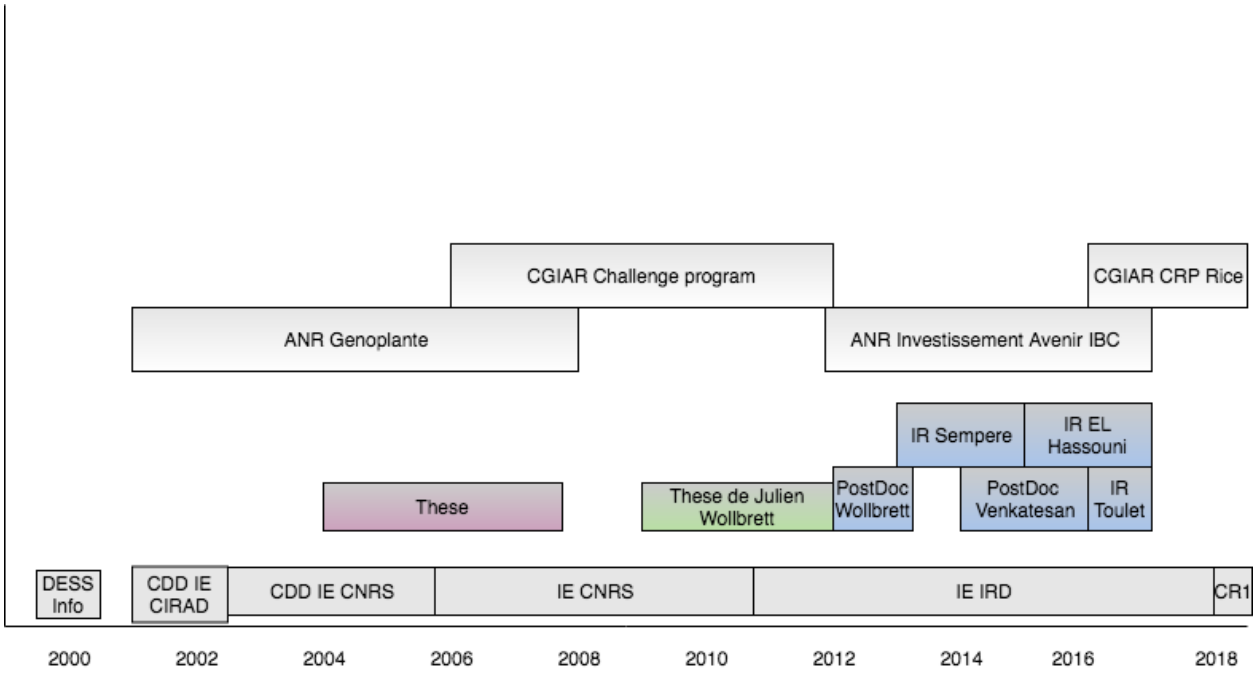


FIGURE 4.1 – Schéma du parcours scientifique

Mon parcours scientifique (cf. schéma 4.1) 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 T-DNA chez la variété Oryza sativa sous-espèce nipponbarre. Le projet était très ambitieux sur le plan informatique et comprenait tous les aspects de traitement de séquences génomiques mais aussi l'informatisation des processus d'analyse phénotypique (ce que l'on appelle aujourd'hui le phénome). Un défi 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 concerné 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. Montpellier) et Manuel Ruiz (Chercheur Cirad). D'un point de vue méthodologique, j'ai défini 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) avec un ensemble de vues sur les schémas des sources de données qui tient lieu de schéma global. J'ai obtenu un poste d'ingénieur d'études CNRS dans l'équipe "Intégration des Données" (ID) dirigée par M. Ruiz au CIRAD (UMR DAP) quelques temps avant de soutenir ma thèse.

J'ai poursuivi dans cette voie autour des méthodes d'intégration de données dans le domaine agronomique et je me suis par ailleurs impliqué dans l'animation de la plate-forme bioinformatique SouthGreen.
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).

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 dans la structuration d'une plate-forme bioinformatique naissante transversale 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 4.2.
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 4.2.
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). Récemment fin 2017, j'ai pris la responsabilité du laboratoire informatique en co-direction avec un chercheur Vietnamien. Le laboratoire compte neuf jeunes enseignants chercheurs Vietnamien 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.
J'ai récemment obtenu un poste de chargé de recherche à l'IRD en présentant les thématiques que je décris dans ce mémoire.

4.2 Activités de recherche

Les progrès de la génomique et des outils de phénotypage à haut débit offrent une occasion unique de découvrir de nouveaux gènes. Les ressources génétiques peuvent maintenant être séquencées à faible coût pour étudier de manière fine leur diversité génétique. Les systèmes d'information actuels permettent de plus en plus l'intégration de données hétérogènes, cependant de nombreux challenges existent encore lorsqu'il s'agit d'exploiter le croisement de ces données d'autant plus que ces données sont massives et multi-échelles. Ceux-ci résident dans les traitements de l'information sous-jacente afin de découvrir rapidement des relations gène-phénotype et leur dépendance à l'environnement qui détermine le rendement des cultures dans des environnements divers. Dans cette partie, seront donc présentés de manière concise les premières approches tentant de résoudre l'interopérabilité des base de données (BD) génomiques (sur le plan syntaxique et sémantique), puis celles permettant l'enrichissement sémantique et l'automatisation du requêtage des systèmes intégrés, enfin celles traitant de données massives et de leur sémantique.

Premières approches pour l'interopérabilité des bases de données biologiques

Dans le contexte du projet ANR Génoplante, une collection de mutants T-DNA de riz a été créée puis caractérisée dans l'objectif d'étudier les fonctions de la totalité des gènes de riz. Son étude comprenait le séquençage des gènes mutés par le T-DNA (et d'autres éléments génomiques mobiles) ainsi que la caractérisation phénotypique des lignées de mutants sur plusieurs sites géographiques. J'ai eu la charge du volet bio-informatique. J'ai ainsi développé un workflow de détection et d'annotation des gènes disruptés par le T-DNA nommés *Flanquing Sequence Tag (FST)* (2004-02). J'ai également été impliqué, dans le développement de l'application OrygenesDB [2], une application Web permettant de stocker des données relatives aux séquences générées lors du projet et également les 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 [3], un système d'information permettant d'intégrer et centraliser ces différentes ressources afin de fournir un portail Web unique aux scientifiques (2008-01).

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

1. https://sites.google.com/site/lmiricevn
2. http://orygenesdb.cirad.fr
3. http://oryzatagline.cirad.fr

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'infrastructures permettant la communication d'applications sur des réseaux distribués. L'objectif de mon travail était de permettre 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 développant deux approches basées l'une sur une architecture de médiation et l'autre sur une architecture orienté services (SOA).

Adaptation de *Le Select* pour la médiation de ressources végétales.

La première approche proposait l'intégration de sources à travers l'adaptation d'un système de médiation de données : *Le Select* [114]. Successeur de DISCO [181], *Le Select* utilisait un modèle pivot relationnel proche du SQL, afin d'intégrer de manière transparente les sources de données hétérogènes et distribuées. Le fait que *Le Select* utilise le standard SQL, lui permettait d'interagir avec un bon nombre d'applications s'appuyant sur ce standard. Les données ont une représentation uniforme exprimée dans le modèle de données relationnel étendu à des types de données définis par l'utilisateur (e.g. structurés, semi-structurés, etc). Ecrit en Java, le médiateur propose 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'attacher une documentation structurée sur ces dernières. D'un point de vue réseau, *Le Select* avait une architecture distribuée de type médiateur/adaptateur, 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).

Dans un premier temps, mes travaux ont consisté à l'intégration syntaxique de plusieurs sources de données à l'aide de *Le Select* par la mise en place d'adaptateurs. Plusieurs sources étaient identifiées dont celles du CIRAD (incluant Orya Tag Line et OrygenesDB), une à l'IRD, une au CNRS (Univ. Perpignan) et une banque d'image 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.). 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) pouvaient être utilisés à cette fin. Ainsi, des règles ACI (correspondance inter-schéma) ont été développées pour chaque adaptateur au niveau des tables et des attributs. La traduction des éléments en ACI a été réalisée selon des règles établies par le document de spécification ODM (Métamodèle Définition Ontologie) [4]. Un schéma global sous la forme d'une vue a été construit sur la base des schémas exposés par les adaptateurs des instances Le Select. Les règles ACI ont été prises en compte dans la construction. Finalement, une ontologie en OWL DL a été développée sur la base du schéma global dans l'objectif de constituer un support au mapping pour l'intégration de nouvelles sources. Enfin, afin

4. https://www.omg.org/spec/ODM/About-ODM

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.

Intégration de données par le biais de Service Web.

La deuxième approche proposait l'intégration des sources à travers l'enchaînement de services Web (SW) grâce à un environnement Web personnalisé (2008-02). Ce système utilisait le Framework BioMoby [196, 195] et son annuaire de services Web bioinformatiques utilisant le protocole SOAP (Simple Object Access Protocol). BioMoby est un projet open source essentiellement orienté sur la découverte et l'exécution de SW biologiques. En effet, par le biais d'un annuaire central, l'application propose 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 permet de faciliter la recherche et l'enchaînement des services. Toutefois BioMoby ne propose pas d'outils permettant de gérer les conflits de noms, lors de l'enregistrement des services. Par ailleurs, il ne propose pas non plus d'outils d'enchaînements de SW intégrés à son API.
Étant donné que BioMoby était très utilisé dans la communauté bioinformatique, la gestion des enregistrements de services et l'orchestration de services étaient devenu deux verrous importants. Mes premiers travaux ont porté sur la réalisation d'un système d'enchaînement de services BioMoby (2007-01 et 2008-02). Les méthodes développées ont permis d'exposer les systèmes d'information développés au cours de ma thèse (i.e. Oryza Tag Line et OrygenesDB) avec des Services Web BioMoby. Des méthodes ont été développées pour orchestrer ces SW avec d'autres services BioMoby. Ces méthodes comprenaient des composants permettant de déterminer i) le type de données (ou d'objet) utilisé par le service, ii) rechercher les services compatibles, iii) déterminer les étapes pouvant être exécutées en parallèle, iv) sérialiser les résultats dans divers formats. Par ailleurs, une application Web a été développée afin de permettre l'utilisation de ces méthodes par un public de biologistes.

Les contributions développées au cours de ma thèse, ont permis de généraliser l'utilisation des standards d'échanges de données (au sein du CIRAD et de ses partenaires) ainsi que d'appliquer différentes approches d'intégration de données en agronomie. Elles ont 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

Génération automatique de services Web pour exploiter des bases de données relationnelles biologiques

Ayant rejoint l'équipe intégration des 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 générées 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, GCP Domain Model [22, 190]. Le modèle GCP a également été utilisé pour implémenter le Framework BioMoby qui s'était rapidement imposé comme protocole standard d'échanges de données et de services dans le domaine bioinformatique.

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 dépasser l'hétérogénéité sémantique de sources de données inter-connectées. 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 [7, 82]. Parmi les technologies utilisées pour exposer des données sur le Web de données RDF, RDFS, OWL et SPARQL sont les éléments importants.

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. L'intégration de base de données relationnelles (BDR) en utilisant les standards du Web Sémantique est confrontée à la problématique de mise en correspondance (mapping) entre des schémas de BDR et une ou plusieurs ontologies. 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 [18, 19] et par la spécification récente d'un langage de mapping nommé R2RML [171] 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) [119].

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 plus d'une ontologie. 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.

Dans notre approche nous avons décidé de détourner D2RQ pour, en plus d'homogénéiser des schémas hétérogènes, 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. 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 [200] capable de prendre en compte les particularités des schémas relationnels. Nous avons ainsi, développé BioSemantic, une approche flexible, générique et automatisée en nous appuyant sur des standards du Web Sémantique et des Services Web (2013-01).

BioSemantic propose la création d'adaptateurs en deux étapes distinctes. Une première étape consiste à créer automatiquement une vue RDF du schéma de la BDR à intégrer, puis à annoter manuellement cette vue à l'aide de termes ontologiques. L'étape d'annotation est la seule nécessitant un utilisateur expert ayant une connaissance du schéma de la BDR à intégrer. La 2ème étape est l'étape de création d'adaptateurs à proprement parler. Elle utilise toutes les vues RDF précédemment créées et annotées pour créer automatiquement des adaptateurs. Dans cette seconde étape aucune connaissance des schémas de BDR n'est nécessaire. La seule nécessité est la connaissance des termes ontologiques utilisés dans le schéma global. La création de ces adaptateurs est basée à la fois sur un enrichissement sémantique de la vue RDF créée par D2RQ et sur la notion de parcours de graphe.

L'utilisation de D2RQ dans BioSemantic présente plusieurs avantages comme la présence d'un langage déclaratif permettant de définir des mappings complexes entre le schéma de la BDR et des termes ontologiques. Un autre avantage est la présence du moteur D2RQ permettant de transformer automatiquement une requête SPARQL interrogeant une vue RDF en une requête SQL interrogeant la BDR sans exporter ses données. De plus, l'utilisation de RDF et son formalisme en triplet va nous permettre de parcourir la vue RDF comme un graphe et ainsi virtuellement parcourir le schéma de la BDR pour trouver une requête pertinente à intégrer dans notre adaptateur (une requête créée automatiquement sera considérée comme pertinente si les données qu'elle renvoie sont identiques aux données renvoyées par une requête créée manuellement par un expert du schéma de la BDR.). Nous allons donc utiliser D2RQ tout en détournant son utilisation pour automatiser la création de requêtes SPARQL.

Enrichissement sémantique d'une vue RDF D2RQ

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) par rapport à un simple graphe RDF, est la présence de relations bien définies entre les tables. Ce sont ces relations qui enrichirons la vue RDF.

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 [141]. 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 bi-directionnelles 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 de l'agrégat 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.

Cependant, 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.

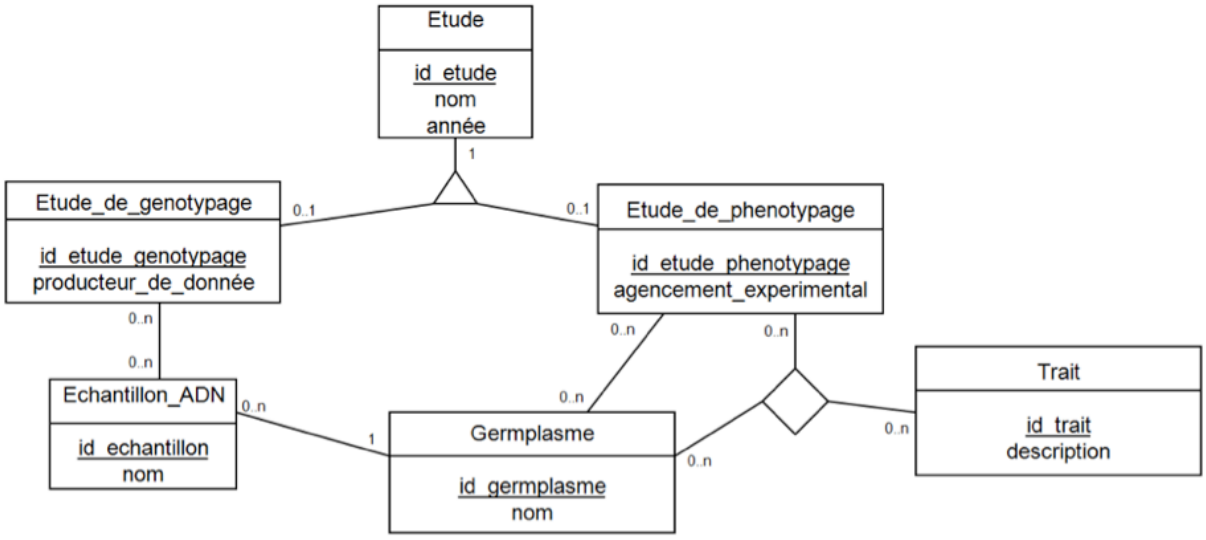


FIGURE 4.2 – Schéma conceptuel montrant une relation d'héritage.
Les entités Etude_de_genotypage et Etude_de_phenotypage sont des spécialisations de l'entité Etude

Passage au modèle relationnel. Les entités du schéma conceptuel donnent lieu à des tables de la BDR. Les relations binaires classiques (d'arité 2), agrégats, compositions ne donnent pas naissance à une table si les cardinalités sont de type (0,n) (1) ; si les cardinalités sont de type (0,n), (0,n) elles donnent naissance à une table d'association.

La conversion d'une relation d'héritage peut s'effectuer de trois façons différentes lors du passage du modèle conceptuel vers le modèle relationnel : soit en aplatissant vers le haut, soit en aplatissant vers le bas, soit en n'aplatissant pas [50]. Les relations d'héritage aplaties vers le haut et vers le bas ne posent pas de problème de combinaison de chemin lors de notre recherche du plus court chemin. En suivant une approche non aplatie, chaque classe est convertie en un schéma de relation. Une clé étrangère supplémentaire, correspondant à la clé primaire du schéma de la relation généraliste, est présente dans chaque schéma de relation spécialisée.

Détection des relations d'héritage non aplaties. La détection automatique de relations d'héritage non aplati pour la transformation d'un schéma relationnel vers une ontologie a été décrite dans [179]. Elle est également utilisée pour typer les relations d'héritage de l'outil DB2OWL [39]. Cette détection automatique est basée sur les techniques de rétro-conception dans les BDR, tentant notamment de convertir un modèle relationnel en modèle entité association [44]. Cette détection utilise la particularité des contraintes d'intégrité entre les tables généralistes et les tables spécialisées. 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. 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 [160]. 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)
avec
Rel(r)  r est une relation
PK(x,r) x est la clé primaire de r
FK(x,r,y,s) x est la clé primaire de la relation r
        et référence y dans la relation s
```

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

Elle détecte également les relations d'agrégation ou de composition. Cela nous permettra ultérieurement donc de détecter tous les types de chemins que nous souhaiterions combiner lors de la création de nos requêtes.

Vue RDF enrichie. 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 4.3.

```
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 d'associations puis les arités des relations. 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 4.3. 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.

Génération de requêtes avec sélection du plus court chemin

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 d'un schéma de BDR. Dans l'exemple de la Figure 4.3, nous allons sélectionner les annotations sémantiques *gcpdm :etude* et *gcpdm :germplasm* pour créer automatiquement une requête renvoyant tous les germplasms d'une étude donnée.

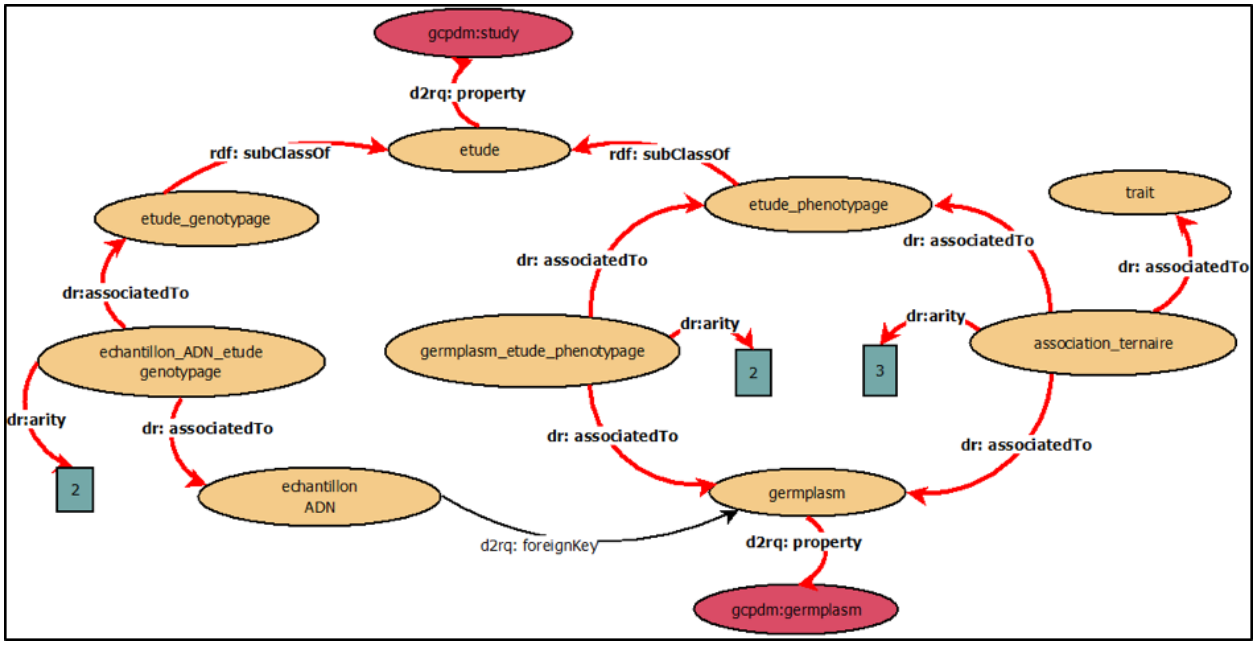


FIGURE 4.3 – Graphe représentant la vue RDF du schéma de la base de données utilisée comme exemple de la création de requête SPARQL.

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 4.4). 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 4.4, l'algorithme passera par la flèche rouge de gauche et ne parcourra 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 4.4.

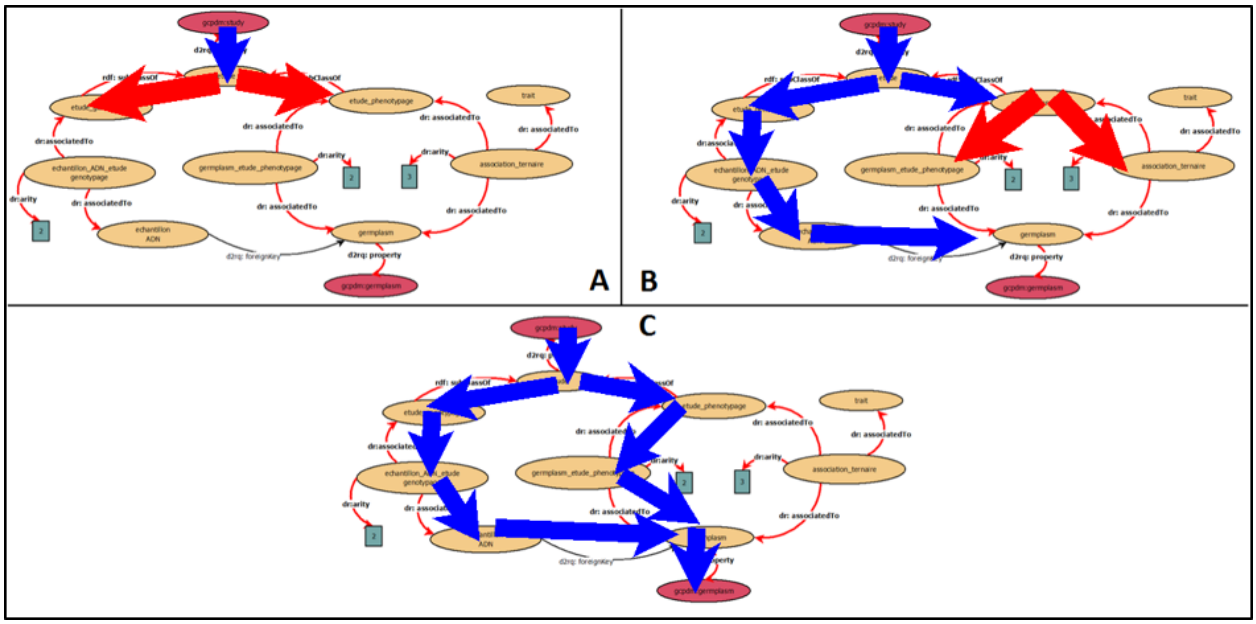


FIGURE 4.4 – Utilisation de l'enrichissement sémantique dans le parcours de graphe.

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

Passage à l'échelle dans l'analyse des variations génomiques

Les enjeux du stockage et traitement des données génomiques sont au cœur des problématiques de l'unité DIADE IRD. 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). Aujourd'hui bien que les biologistes aient souvent encore l'habitude de manipuler leurs données sur leur poste de travail à l'aide de logiciels de type tableur, les données devenant massives et complexes leurs en limite l'utilisation. Il en résulte que les alternatives souvent proposées jusqu'alors passaient par l'utilisation de traitements automatiques exécutables en lignes de commandes ou avaient recours à des SGBD de type relationnels qui passaient difficilement à l'échelle dans certains cas. Avec mes collègues, nous souhaitions lever ces verrous en proposant une nouvelle approche utilisant les SGBD NoSQL.

Gèrer le stockage de la masse de données.

Depuis 2013, je suis impliqué dans le développement d'un logiciel nommée Gigwa 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, standard de représentation des variants 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 ces sous-jeux de données sous divers formats standardisés 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 leader dans ce domaine. Un article effectuant le comparatif et décrivant l'application a été publié en 2016 (2016-01) [159].

Actuellement, nous développons une API de services REST pour Gigwa comme alternative à son interface web. Cette API qui respecte et étend les recommandations du GA4GH Data Working Group [5] et de la Breeding API [6] permettra d'accroître l'interopérabilité de l'application avec d'autres systèmes utilisés dans la communauté bioinformatique tels que Galaxy [64, 65], FlapJack [121], SniPlay [46] ou Toggle [123].

De la masse de données à la connaissance.

En ayant en perspective l'idée d'extraire de la connaissance de cette masse d'information génomique, nous avons commencé à explorer les possibilités que peuvent offrir les approches de « mapping » entre le Web sémantique et les bases de données NoSQL orientées documents (e.g., MongoDB). Dans le cadre du projet spirale IRD BIOeSAI, 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. L'aspect novateur de ce projet réside dans la mise au point d'un système évolutif permettant aux utilisateurs d'effectuer toutes les étapes allant de l'intégration de données jusqu'à la composition de requêtes. Ce système inclut également la gestion des métadonnées et des annotations ajoutées par les utilisateurs. Toutefois, la méthode mise en place ne permet pas de détecter des relations explicites/implicites entre les données gérées par le système. Par exemple, il n'est pas possible de déduire qu'une région géographique (localisation GPS ou décrite) est incluse dans une région plus large afin d'agréger des résultats. Ou encore, il est impossible de propager une information qui est implicite comme "une maladie affectant une la plante affectera tous ses tissus.

Au cours de son stage de M2, Luyen Le Ngoc évalua plusieurs approches pour mapper le schéma du système (Document JSON) à un modèle RDF annoté avec des ontologies biologiques [111] (2016-02). Il aborda notamment les approches de matérialisation de données en triplets RDF avec xR2RML [120] et de ré-écriture de requêtes avec les applications Ontop [151], xR2RML et AllegroGraph. Puis dans une autre optique de matérialisation, il évalua le SGBD 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 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.

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

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 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) [130]. Elles comportaient à la fois des données textuelles et des images avec métadonnées 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 trés bon résultats sur l'ensemble des tests. Virtuoso édition libre, a obtenu de bons résultats pour les logiciels libres. Ces résultats publiés (2016-02) [111], nous ont confortés dans l'utilisation de Virtuoso pour la suite de nos travaux.

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
— (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)

Vers de nouvelles approches d'intégration sémantique des données agronomiques

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 [34], développé par l'équipe INRIA Zénith, Bioportal [134, 118] et Biosemantic [200] en réalisant un prototype [27]. Ce travail a permis d'identifier un besoin dans l'annotation sémantique des données et la gestion des ontologies pour le domaine des plantes.

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

AgroPortal [83, 81] 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

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

concerné. Le portail propose des services de recherche d'ontologie 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 [118] ont été développées pour classer les mappings avec des termes ontologiques. Un nouvel algorithme de recommandation a été implémenté dans le *Recommender* [152] et un nouveau modèle de métadonnées a été développé et implémenté dans la plateforme [182].

La plateforme AgroLD. Ainsi, une infrastructure capable de gérer des ontologies du domaine agronomique et de proposer des services pour rechercher, annoter les données était dorénavant disponible. Toutefois, les données ainsi annotées sémantiquement nécessitaient une infrastructure permettant de les gérer efficacement. Or, il existait des projets équivalents dans le domaine biomédical et bioinformatique Bio2RDF [14, 26], EBI RDF [85], ou encore Uniprot RDF [145] mais pas encore de projet dans le domaine agronomique. Avec l'aide d'un post-doctorant recruté dans le projet IBC en 2014, nous avons élaboré des modèles de données et développé un premier prototype de système d'intégration sémantique de ressources biologiques : AgroLD [9]. AgroLD 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. Le cadre conceptuel de la connaissance est basé sur des ontologies bien établies dans le domaine telles que Gene Ontology [9, 178], Plant Ontology [142], Plant Trait Ontology [36], Plant Environement Ontology [24] pour n'en citer que quelques unes dont la majorité sont hébergées par le projet OBO Foundry [169]. En outre, compte tenu de la portée de l'effort, nous avons décidé de construire AgroLD en plusieurs phases. La phase actuelle (première) 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 [177] , UniProtKB [113] , Gene Ontology Annotation [11] ainsi que des ressources développées par la plateforme SouthGreen [10] comme TropGeneDB [71], OryGenesDB [48] , GreenPhylDB [153] , OryzaTagLine [99] , SniPlay [47]. Le tableau 4.1 donne un appercu des espèces et sources intégrées.

Nos contributions portent sur la création de différents workflows 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], aucun n'étaient adaptés pour prendre en compte la complexité des formats de fichiers plats du domaine biologique ou même la complexité des informations qu'ils pouvaient contenir. Déjà, initié avec le projet BioSemantic, nous avons étendus ces modèles de transformation à une plus large palette de standards de données en génomique et phénomique tels que le Generic Feature Format (GFF) [13], le Gene Ontology Annotation File (GAF) [14], le Variant Call Format (VCF) [42], le Genomic Feature

9. http://www.agrold.org
10. http://southgreen.fr/
11. https://project.inria.fr/datalift
12. http://purl.org/twc/id/software/csv2rdf4lod
13. http://gmod.org/wiki/GFF3
14. http://geneontology.org/page/go-annotation-file-format-20

and Variation Ontology(GFVO) [10] et le Minimum Information About a Plant Phenotyping Experiment (MIAPPE)) [90] et travaillons actuellement à packager ces modèles dans une API [15].

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 fournies par la source d'origine. À la fin de la phase 1, début 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 http://purl.obolibrary.org/obo/GO_0005524). 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 annotation 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 http://purl.obolibrary.org/obo/PO_0025034) ou un caractère phénotypique (plant height serait annoté avec le concept ayant pour URI http://purl.obolibrary.org/obo/TO_0000207). Comme le montre la figure 4.5, 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) [16] et mieux contrôler l'annotation sémantique faite par AgroPortal et y gérer les différentes annotations pour un résultat optimal.

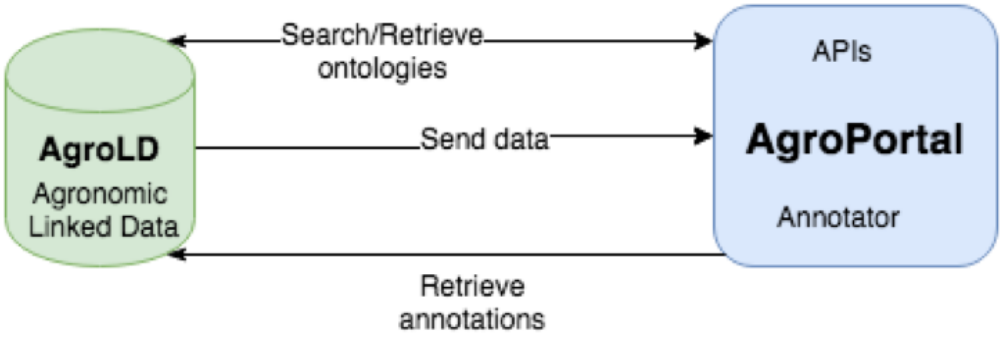


FIGURE 4.5 – Processus d'annotation sémantique entre AgroPortal et AgroLD

15. https://github.com/pierrelarmande/AgroLD-ETL
16. https://github.com/pierrelarmande/ontology-project

Sources de données	**URLs**	**Format de fichier**	**Nb Tuples**	**Especes**	**Ontologies utilisées**	**Nb de triplets produits**
Oryzabase	shigen.nig.ac.jp/	Custom flat file	17K	R	GO,PO,TO	153K
GO associations	geneontology.org	GAF	1, 160K	R, W, A, M, S	GO	2, 700K
OryGenesDB	orygenesdb.cirad.fr	GFF	1, 100K	R, S, A,	GO, SO	2, 300K
Gramene	gramene.org	Custom flat file	1, 718K	R, W, M, A, S	GO, PO, TO, EO	5, 172K
UniprotKB	uniprot.org	Custom flat file	1, 400K	R, W, A, M, S	GO, PO	50, 000 K
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
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 ont annoté 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 (caracteres spécifiques des plantes). 134 ontologies)

Les graphes RDF sont nommés d'après les sources de données correspondantes, partageant un espace de noms commun : http://www.southgreen.fr/agrold/. Les entités dans les graphes RDF sont liées par des URI communs partagés. Comme principe de conception, nous avons utilisé des schémas d'URI mis à disposition par les sources (par exemple, UniprotKB) ou par le registre Identifiers.org [97]. Par exemple, les protéines de UnitProtKB sont identifiées par l'URI de base : http://purl.uniprot.org/uniprot/ ; les gènes intégrés à partir de Gramene / Ensembl plant sont identifiés par l'URI de base : http://identifiers.org/ensembl.plant/. Lorsqu'elles ne sont pas fournies par les sources ou par Identifiers.org, de nouvelles URI ont été construites tels que TropGene et OryGenesDB ; dans ce cas, les URI prennent la forme http://www.southgreen.fr/agrold/[resource_namespace]/[identifier]. De plus, les propriétés reliant les entités se présentent sous la forme : http://www.southgreen.fr/agrold/vocabulary/[property]. À propos de la liaison d'entité, nous avons utilisé «l'approche basée sur la clé» qui est la plus courante. Elle combine l'identifiant unique de l'entité partagée avec la communauté plus le modèle de base d'URI de la ressource. De plus, nous avons également respecté «l'approche URI commune» qui recommande d'utiliser le même modèle d'URI lorsque le même numéro identifiant est utilisé dans différents jeux de données. Par conséquent, définir le même URI pour des entités identiques (représentées par des identifiants) dans différents jeux de données permet d'agréger des informations supplémentaires pour cette entité. De plus, nous avons utilisé des liens de références croisées (représentés par des identifiants) en les transformant en URI et en reliant la ressource au prédicat *rdfs :seeAlso* ou *has_dbxref* si la référence n'a pas d'URI . Cela augmente considérablement le nombre de liens sortants, rendant AgroLD plus intégré avec d'autres sources de données. À l'avenir, nous comptons mettre en œuvre une «approche basée sur la similarité» 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 utiliser 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 léger [17] qui associe les classes et propriétés identifiées dans AgroLD avec des ontologies correspondantes. Par exemple, la classe Protein (http://www.southgreen.fr/agrold/resource/Protein) est associée à la polypeptide (http://purl.obolibrary.org/obo/SO_0000104) 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 http://www.southgreen.fr/agrold/vocabulary/has_function, 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* (http://www.southgreen.fr/agrold/vocabulary/has_trait), relie les entités aux termes TO équivalent. Elle est associée à une propriété plus générique. Pour l'instant, 55 mappings ont été identifiés. De plus, les mappings sont stockés avec les ontologies dans AgroPortal, ce qui permet des liens directs entre les classes et les instances de ces classes dans AgroLD. Par ailleurs, les classes, les propriétés et les ressources sont dé-référencées sur un serveur Pubby [40] dédié (par exemple, http://www.southgreen.fr/agrold/page/biocyc.pathway/CALVIN-PWY).

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 (e.g., bioinformaticiens, 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).

17. https://github.com/SouthGreenPlatform/AgroLD

Ainsi la plateforme AgroLD fournit 4 points d'entrée :

— **Quick Search** [18], un plugin de recherche à facettes mis à disposition par Virtuoso, qui permet aux utilisateurs d'effectuer des recherche par mots-clés et de parcourir le contenu d'AgroLD en naviguant dans les liens ;

— **SPARQL Editor** [19], un éditeur de requêtes SPARQL qui fournit un environnement interactif pour la formulation de requêtes SPARQL. Avec un étudiant de M2, nous avons développé l'éditeur en se basant sur les outils YASQE et YASR [147] et l'avons adapté pour notre système.
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)* en utilisant Quick Search, renverrait 73 résultats alors qu'utiliser un *property path* dans une requête SPARQL renverrait 137 résultats. Par ailleurs, le langage SPARQL étant plus expressif il est possible de composer des requêtes complexes recherchant sur plusieurs graphes. Toutefois, parce-que 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 ajouté 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** [20], est une implémention de RelFinder [75] qui permet aux utilisateurs d'explorer et de visualiser les relations existantes entre entités. Les relations entre des objets constituent une information importante. Ce mode de recherche permet de partir d'un point de départ et explorer le graphe, ou éliminer des parties du graphe des données à l'aide de filtres pour découvrir des relations. Ces techniques présentent malheureusement l'inconvénient de demander beaucoup d'effort manuel de la part de l'utilisateur. L'application RelFinder découvre automatiquement de telles relations dans une source de données disposant d'un serveur d'accès SPARQL et les affiche sous forme de graphe. Elle aide l'utilisateur à trouver les entités avec une fonctionnalité d'auto-complétion associée à une gestion des cas d'ambiguïté où les résultats sont classés par ordre de pertinence (entités ayant le plus de label contenant l'expression saisie). 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 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 ;

— **Advanced Search** [21], un formulaire proposant des recherches spécifiques par entité et possédant un moteur d'agrégation de ressources externes. Le formulaire Advanced Search est basé sur une API REST [22], entièrement développée dans le cadre du projet AgroLD. Le but de ce formulaire est de fournir aux utilisateurs non-techniques 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

18. http://www.agrold.org/quicksearch.jsp
19. http://www.agrold.org/sparqleditor.jsp
20. http://www.agrold.org/relfinder.jsp
21. http://www.agrold.org/advancedSearch.jsp
22. http://www.agrold.org/api-doc.jsp

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)

Dans le domaine de la Recherche d'Information, nous avons commencé à travailler sur l'indexation des graphes RDF et leur enrichissement sémantique (2016-04). Afin d'améliorer la fonctionnalité Quick Search en récupérant plus d'informations textuelles et en masquant les détails techniques, nous avons développé un outil d'indexation permettant de communiquer facilement entre Virtuoso et des clusters Elastic. Ainsi, cet outil permet d'indexer les fichiers Json et de gérer les index (par exemple, mettre à jour, supprimer) sur des clusters Elastic sans utiliser cURL. La difficulté résidait dans le passage d'une structure de donnée sous forme de graphe dans un document document JSON où les relations sont par nature aplatis. Par exemple, des fonctionnalités de chaînage des propriétés ont été développées afin de récupérer l'information textuelle associée aux propriétés des triplets. Par ailleurs, l'indexation des URI et identifiants de bases de données externes ont été rendu plus explicites en utilisant des fonctions de recherche http pour récupérer de l'information supplémentaire. De plus, nous avons développé un outil d'annotation générique qui facilite la communication avec NCBO Annotator pour annoter des fichiers JSON avec des ontologies disponibles à partir d'AgroPortal [81]. Nous utilisons cet outil avec AgroPortal Annotator pour enrichir et indexer les fichiers Json avec des informations supplémentaires à partir de termes ontologiques tels que des labels, des synonymes, des termes parents et enfants, etc.

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 [77, 110]. 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 intéressant d'évaluer, d'exploiter, voire d'étendre. Nous avons développé un test de référence (*Gold Standard* [23]) 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 [77, 110, 122, 129] pour mettre en place une classification des systèmes SQR en fonction des principales approches explicitées précédemment, illustrées à la figure 4.6, et effectué une évaluation empirique de ces systèmes. Finalement, nous avons porté notre attention sur le système LODQA [89]. Le système est basé sur 3 modules i) décomposition de la phrase en langage naturel, ii) un module de correspondance de termes, iii) réécriture en requête SPARQL. LODQA étant très dépendant du domaine, dont les mots 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.

Sélection de références

— (2018-01a) 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

23. Gold Standard : est 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.

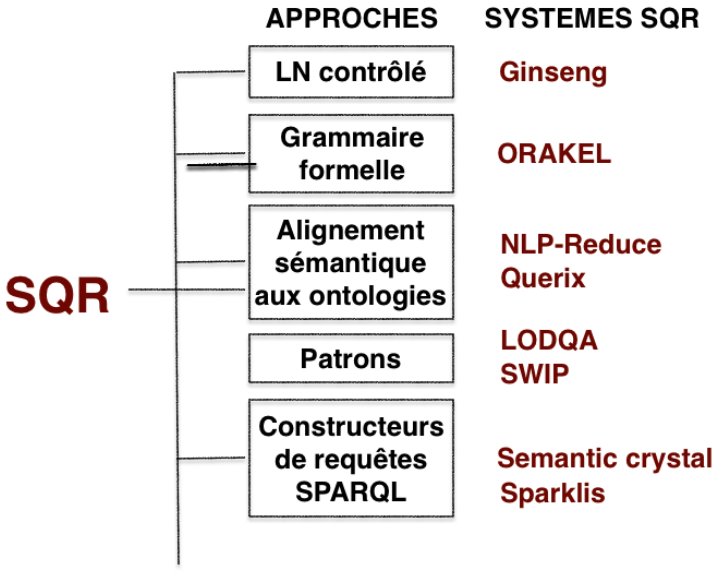


FIGURE 4.6 – Vue générale de notre classification SQR

— (2018-01b) Do H., Than K., and Larmande P. Evaluating Named-Entity Recognition approaches in plant molecular biology. MIWAI 2018.Springer 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)
— (2018-02) Venkatesan A., Tagny G., El Hassouni N., Chentli I., Guignon V., Jonquet C., Ruiz M., and Larmande P. Agronomic Linked Data (AgroLD) : a Knowledgebased System to Enable Integrative Biology in Agronomy. Plos One (In Press) Impact Factor : 2.766
— (2018-01) 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, 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, 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. 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) Imène Chentli, Pierre Larmande et Konstantin Todorov. Construction d'un gold standard pour les données agronomiques. IC2016, Montpellier, France.
— (2016-06) Dagmara Robakowska Hyzorek, Marie Mirouze, Pierre Larmande. 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.

Chapitre 5

Projet

Ce chapitre présente les perspectives ouvertes par les recherches déjà menées. Il est dédié à la présentation des directions et évolutions des recherche pressenties dans les cinq prochaines années ...et bien sur corrélées aux étudiants que je supervise et aux projets auxquels je participe.

5.1 Objectifs

Mon projet de recherche aborde le problème suivant : Comment gérer et structurer 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, l'objectif de ce projet est de construire des réseaux d'interaction moléculaires entre gènes à partir de données éparses (articles scientifiques, bases de données publiques, données expérimentales, ...) afin d'identifier les gènes clés pour l'amélioration des plantes.

Compte tenu de l'ambition du projet, je compte organiser mon travail dans des activités qui seront menées en parallèle sur une durée de 5 ans. Chacune de ces activités propose de traiter (voire de résoudre) des verrous scientifiques relevant des domaines informatique et bioinformatique. Par ailleurs, je compte obtenir des financements sur appels d'offres pour développer ces activités sur un plus long terme.

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 de recherche sont envisagées : intégration dynamique des données, enrichissement des connaissances, priorisation de gènes candidats.

Dans ce processus, 1) une première voie consistera à transformer et intégrer dynamiquement ces données dans AgroLD pour les rendre plus facilement utilisables en terme algorithmique. Une attention particulière sera portée aux données expérimentales produites par les chercheurs des unités plantes Montpelliéraines (DIADE, AGAP, IPME) et les partenaires du sud (LMI RICE, IRRI). S'agissant souvent de données massives (i.e. données de génotypages ou d'images), 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 un portail d'ontologies de référence pour le domaine agronomique. Puis, afin d'enrichir les liens entre les différents graphes générés et ainsi produire un réseau d'interaction qui permettra la découverte

de nouvelles connaissances, de nouvelles méthodes de liage de données RDF spécifiques aux problématiques bioinformatiques seront développées. Par ailleurs, le développement de méthodes de fouille de texte pour identifier les entités biologiques et leurs relations dans les publications scientifiques sera envisagé. Enfin, afin de permettre une recherche d'information efficace et trier pertinemment les résultats, plusieurs méthodes et algorithmes de priorisation de gènes candidats seront évaluées et proposées.

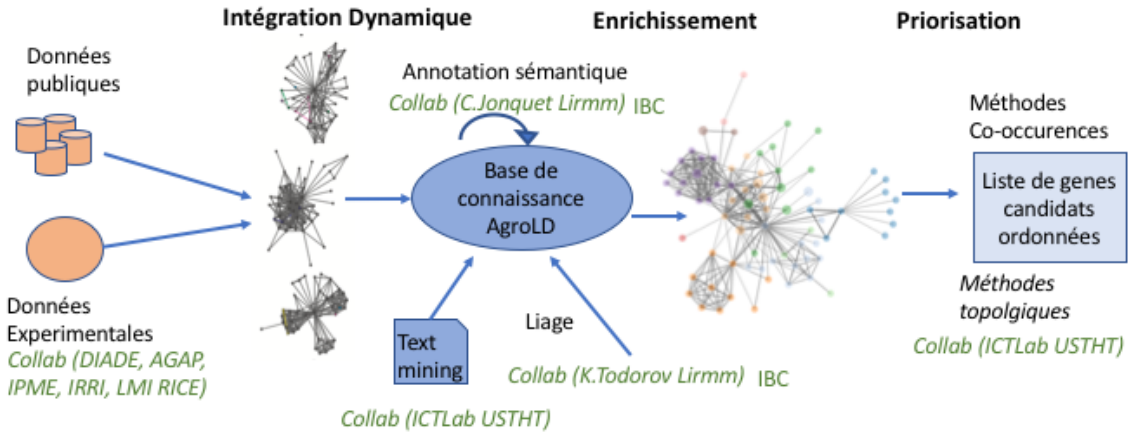


FIGURE 5.1 – Schéma général du projet de recherche

Enfin, afin de permettre une recherche d'information efficace et trier pertinemment les résultats, plusieurs méthodes et algorithmes de priorisation de gènes candidats seront évaluées et proposées.

5.2 Intégration de données et extension de connaissances

5.2.1 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'interaction moléculaire (réseaux de co-expression de gènes, transcriptomique, Facteur de transcriptions, complexe proteine-proteine). Ce processus de transformation est souvent appelé le **lifting de données** [1].

Je m'appuierai sur le projet AgroLD (Agronomic Linked Data) 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 spécifiques (e.g. TropGeneDB, OryzaTagLine, etc.) soit

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

pour des formats de fichier génériques (e.g. GAF, VCF, GFF, etc.).

Toutefois, une attention particulière sera portée aux données expérimentales produites par les chercheurs de l'unité Diade ou les partenaires du sud (LMI RICE, IRRI). S'agissant souvent de données volumineuses (i.e. données de génotypages ou d'images), la méthode doit éviter de transformer intégralement les données afin de garantir de bonnes performances.

Concernant l'extraction d'information contenue dans des bases de données relationnelles, nous avons développé une application BioSemantic, au cours de la thése de Julien Wollbrett. Cette dernière propose une approche flexible et automatisée pour la création de vue RDF en se basant sur D2RQ 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 [200]. Toutefois, il existe de nombreux outils de transformation et language de mapping associés adaptés aux différents types de SGBD et aux modèles de représentation des données. L'article de Michel et al [119] en dresse un inventaire et compare les différentes méthodes. De plus, les auteurs proposent également xR2RML [120] 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, object-oriented, NoSQL). En collaboration avec l'équipe Inria wimmics, je compte développer ces aspects pour extraire les données expérimentales de variations génomiques actuellement stockées dans Gigwa [159].

5.2.2 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, Sequence Ontology, Plant Ontology(PO), Trait Ontolgy (TO), Phenotype quality ontollogy (PATO) et Environnement Ontology (OE). Un lien sémantique (e.g. annotation sémantique) est créé des 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 [21, 164, 5]. Ces annotations sont souvent produites à partir de logiciels bioinformatiques souvent basés sur des recherches de similarité sur les séquences nucleotidiques. L'article de Blake et al, 2013 [20] donne un apercu 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 [53, 138] . De fait, de nombreuses méthodes sont proposées afin de relier des termes (ou concepts) issus de différentes ontologies mais peu proposent des méthodes efficaces dans le cas de traits complexes comme les maladies ou les phénotypes [72]. Dans notre cas, il y a certaines spécificités dont il faut tenir compte :

— Un terme (peut faire référence à plusieurs ontologies de domaine différents (e.g. Dwarfism -> PATO, Tillering -> PO, Tiller angle -> TO),
— Un terme composé peut être annoté à partir de deux ontologies (e.g. wrinkled seed, aborted seed font référence aux ontologies PATO et PO),

— Un terme biologique peut être représenté par son symbole ou son acronyme,
— Un terme biologique peut être polysémique et ambigu, donc difficile à annoter,

Pour répondre à ces défis et 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 [84]. Agroportal [81] 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) [81], nous comptons développer un workflow d'annotation entre AgroPortal et AgroLD basé sur des mesures de similarités, le traitement de texte (voir section 5.3.1) et utilisant les fonctionnalités d'AgroPortal pour réaliser l'association des données avec les concepts ontologiques.

Il arrive également que dans certains cas les différentes ontologies utilisées ne se recouvrent pas. Afin de relier les différents types de données et propriétés de ces ontologies, j'ai initié le développement d'une ontologie AgroLD qui servira de glu entre les classes et les propriétés identifiées dans AgroLD [2].

5.3 Extraction et exploitation de la connaissance

5.3.1 Extraction d'entités biologiques et de relations

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. 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 délai de mises à jour plus long. 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*).

Dans le domaine de l'extraction de connaissances, une tâche importante consiste à identifier et classer par type les entités biologiques, également appelée des entités nommées [3] .

Afin d'en extraire de l'information pertinente, ici les entités Ehd1, Hd3a, Ghd7 et Hd1 et la relation down-regulated, je souhaite évaluer des approches de « text mining » (NLP). notamment les récents développement qui utilisent les méthodes de deep learning avec des modèles LSTM-CRF (Long Short Term Memory model combinés avec Conditional Random Fields) pour détecter les Entités Nommées [69, 13]. L'accent sera également donné sur la constitution d'un corpus de données sur le riz qui pourra servir de modèle d'entraînement pour détecter des Entités et leurs relations dans le texte. Le but de ce travail sera de proposer un module d'extraction d'information et formalisation de connaissances que les experts puissent valider par l'application AgroLD. Des

2. https://github.com/SouthGreenPlatform/AgroLD_ETL/blob/master/model/agrold_schema.rdf
3. named entity recognition (NER)

premiers résultats d'évaluation ont déjà été obtenus.

L'architecture de LSTM-CRF est illustrée à la figure 5.2 [69]. 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'intégration x_1 ; x_2 ; ... ; x_n pour chaque mot. Chaque vecteur d'inclusion 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.

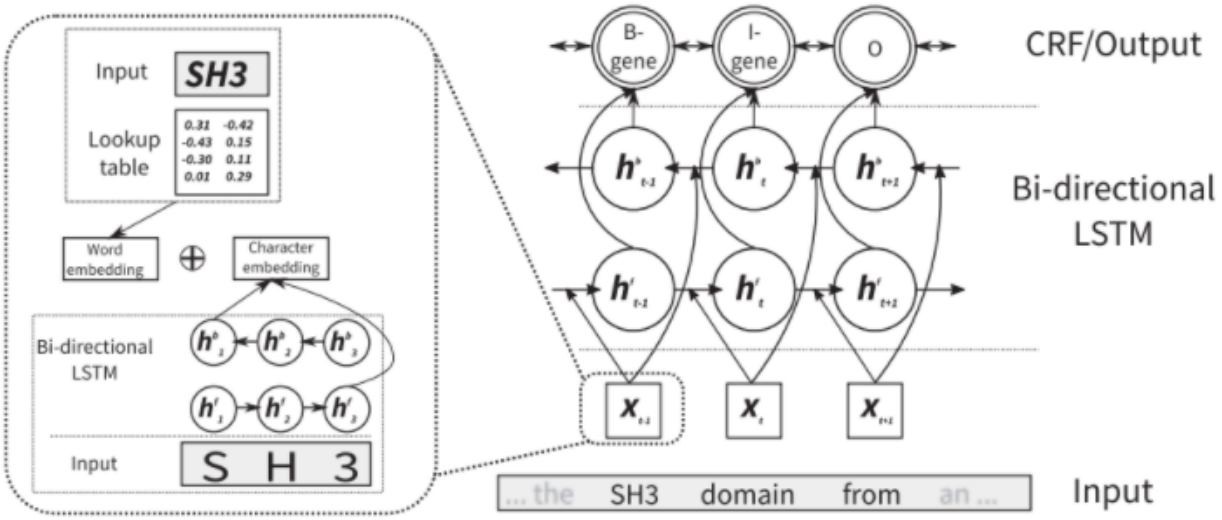


FIGURE 5.2 – Le modèle LSTM-CRF présenté dans [69]

L'approche hybride [13] repose sur un dictionnaire d'entités associé à des classificateurs d'apprentissage automatique. Dans un premier temps le dictionnaire de recherche d'entités 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 accés aux dictionnaires construits sur les bases pubmed du NCBI [4]. Ensuite, le framework Distiller est utilisé pour extraire ces informations en tant que fonctionnalité 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) pour extraire des informations d'un texte. AKE semble être différent de NER, car celui-ci s'intéresse a la recherche du petit ensemble d'information les plus pertinentes 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 [12], telle que le part-of-speech (POS), l'analyse statistique, etc. Il fonctionne avec la possibilité d'implémenter différents pipelines pour

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

différentes tâches. Les modules partagent leur informations sur les entités dans une mémoire partagée afin que les autres modules puissent y accéder. L'implémention 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 5.3.

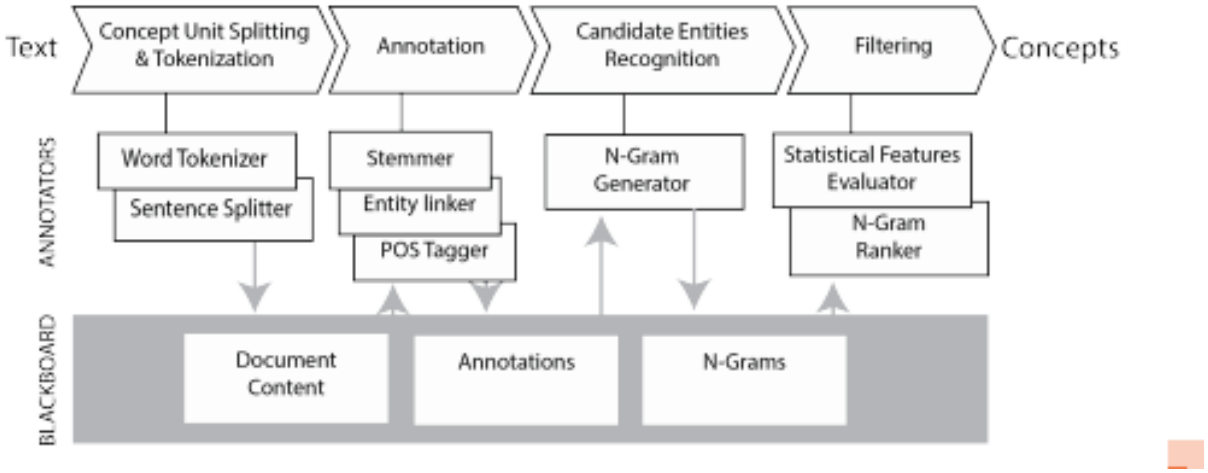


FIGURE 5.3 – Le schéma du Distiller présenté dans [13]

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 résumé, nous avons implémenté 3 méthodes pour évaluer l'approche hybride :

— basée sur les résultats OGER
— 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

Dans ce projet, nous avons utilisé les jeux de données de Oryzabase [5], une base de données intégrée sur le riz. En nous focalisant sur les gènes du riz, nous avons téléchargé les jeux de données *Gene List* et *Reference* contenant respectivement une liste de 21 739 gènes différents connus et un ensemble de résumé d'articles avec des gènes associés. Nous les avons utilisé comme données d'entraînement et de validation, respectivement pour les diverses approches d'extraction d'entités.

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

LSTM-CRF réalise les meilleures performances parmi les modèles. En moyenne, le score F_1 - est égal à 86,72 % pour la méthode générique LSTM-CRF et à 80,44 % pour la méthode générique LSTM.

5. https://shigen.nig.ac.jp/rice/oryzabase/

	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 NER - 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

D'un autre côté, le résultat de la méthode hybride est très exploitable. La performance de OGER avec CRF a donné le meilleur résultat parmi les 3 tests, soit 86,72% en moyenne. Le deuxième correspond à OGER associé au réseau de neurones, qui a atteint une précision de 67% en moyenne. Bien que les améliorations ne soient pas aussi élevées que prévu, il y a eu quelques améliorations, comparé à OGER seul qui se situe autour de 58,5 %. Le résultat est présenté dans le tableau 5.2.

	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 approchesOGER - le résultat des performances de la méthode hybride en termes de précision, de rappel et de $F_1 - score$

5.3.2 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 [55, 2, 91].
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 5.4. 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 de dump 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 paramêtres 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 basés sur l'apprentissage. D'autres outils peuvent éventuellement être utilisés comme d'autres sources de connaissance, par exemple, des dictionnaires de données ou des mappings préalablement définis. La sortie du workflow 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*.

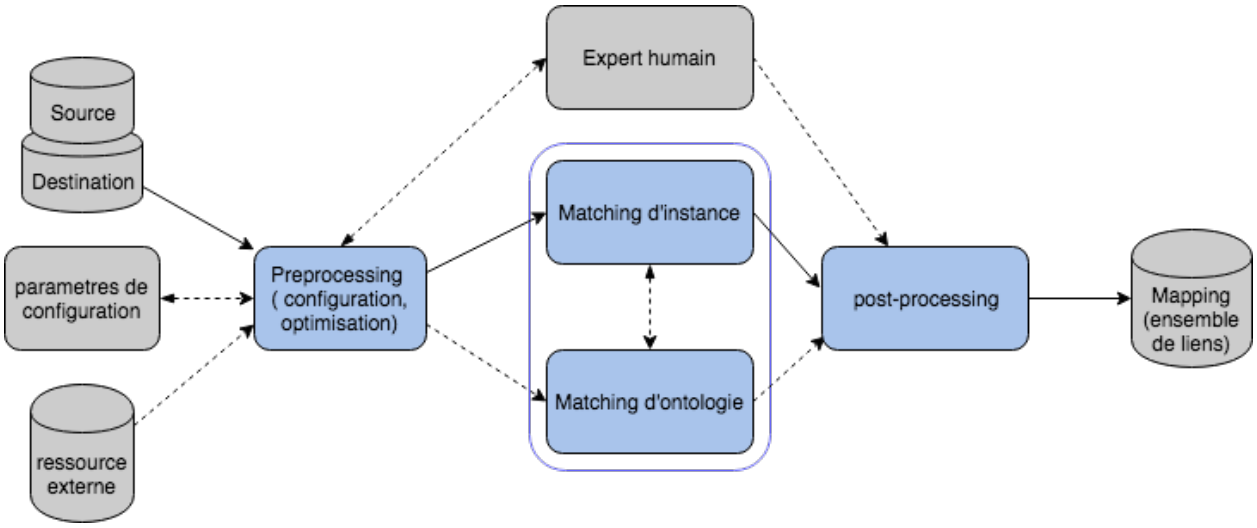


FIGURE 5.4 – Workflow général du processus de liage de données adapté de Achichi et al,2018 [91]

Les challenges du liage de données Il existe de nombreux outils qui tendent à résoudre 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 [80, 131] d'évaluer les relations entre instances. Les jeux de données intégrés dans AgroLD présentent largement ce problème.

L'exemple de la figure 5.5 montre 2 entités 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 : basée sur la valeur, ontologique et logique.

La dimension basée sur la valeur fait référence aux propriétés contenant des valeurs littérales (texte) exprimées en langage naturel ou valeurs numériques et qui peut induire des erreurs de liage. Les auteurs de Achichi et al, 2018 [3] identifient 4 niveaux d'hétérogénéité également indiqués dans la figure : terminologique, linguistique, bonnes pratiques de représentation et sur les types de valeurs.

— *Hétérogénéité terminologique.* 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 abbréviation. Pour pallier ce problème, certaines applications proposent des fonctionnalités d'expansion d'acronymes/abbréviations.

— *Hétérogénéité linguistique.* Les termes concernés sont représentés dans des langages différents. C'est un problème trouvé fréquemment lorsqu'on travaille avec des données expérimentales et qui reflètent la diversité des informations que l'on peut trouver sur le Web. Dans ce cas, 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.

— *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.

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

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

— *Hétérogénéité du vocabulaire.* Les classes et les propriétés sont souvent décrites en utilisant différents vocabulaires par différents producteurs de données, 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 décrit différemment dans les deux entités. Dans ce cas, 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 leur 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). 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.

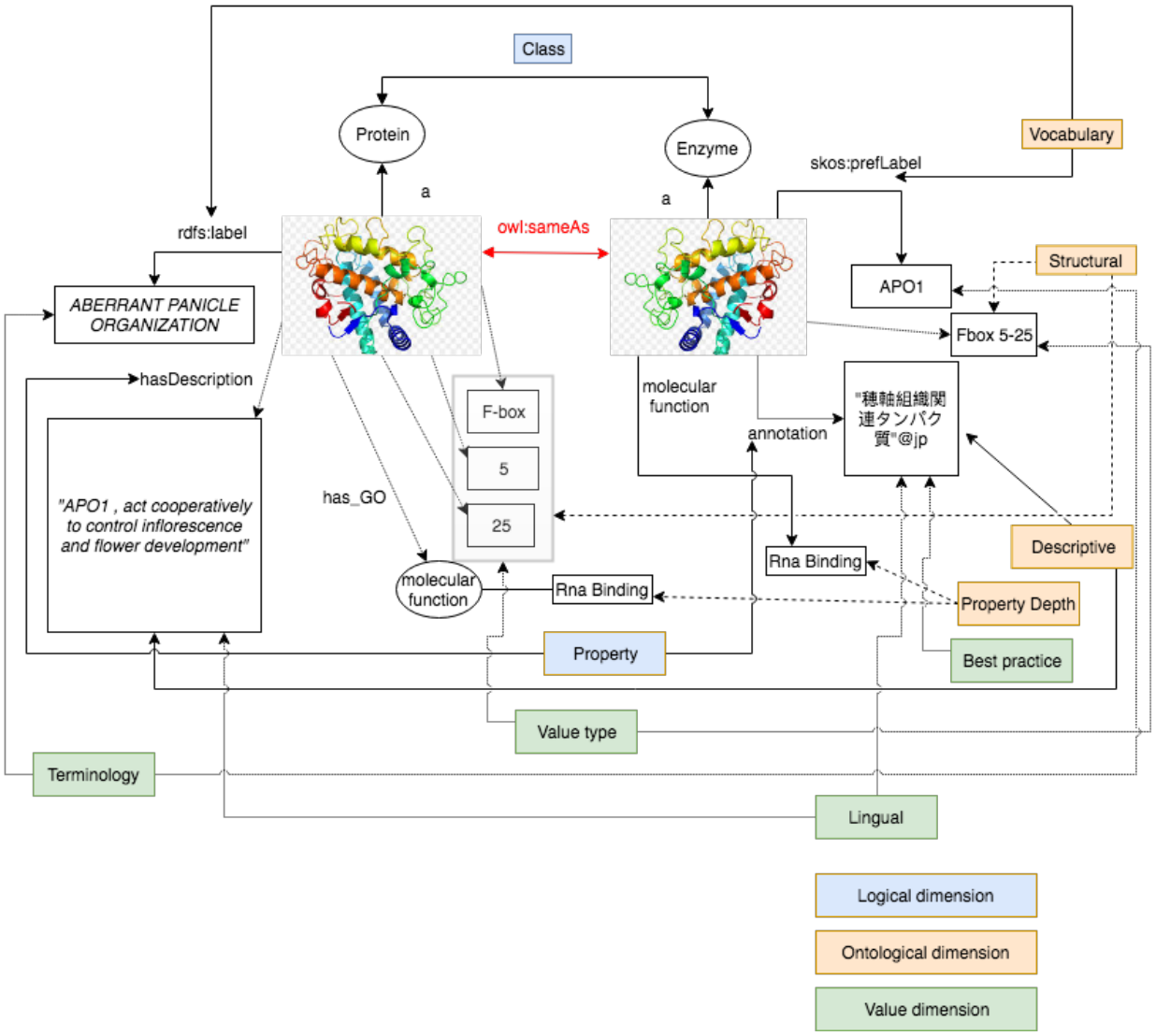


FIGURE 5.5 – Schéma général d'un exemple de liage de données inspiré de [3]

L'état de l'art des méthodes de liage ont été implémentés dans de nombreux logiciels dont les suivants sont les plus cités ou plus récents :

— **Silk** [80] 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** [131] 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'instance) 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** [1] 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 module qui s'enchaîne 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 identifiant d'instances spécifiques à une ressources 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 [173] et un

algorithme de classement de clé [2] sont appliqués sur chaque paire de cluster 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 [6], ce qui permet de garder une trace des indices de confiance associés, ou sous forme de triplet *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 proposerons 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 manières effectives les ressources inter-graphes 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.

5.3.3 Raisonnement sur les données

Le fait que les données soient structurées en RDF présente l'avantage d'utiliser des mécanismes de requêtes tels que SPARQL pour exploiter au mieux les liens explicites existants entre les données (e.g. utilisation des propriétés *is_a* dans les requêtes). Une autre manière d'enrichir ces

6. http://alignapi.gforge.inria.fr/edoal.html

liens est d'utiliser des mécanismes d'inférences qui grâce aux ontologies peuvent déduire de nouvelles connaissances implicites. Pour mener à bien cette problématique j'évaluerai les possibilités que proposent les langages SWRL [7], SPIN [8] et SHACL [9] pour implémenter des règles d'enrichissement ou de vérification de cohérences dans les 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 [37].

5.4 Applications sur les graphes de connaissances

5.4.1 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 [126]. Par exemple, Endeavour [126, 183] a pu associer le gène GATA4 à une hernie diaphragmatique congénitale ; GeneDistiller [157] 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 [23, 125, 45]. Le second type de **méthodes est basé sur la fouille de texte** [51, 166, 167]. En général, ces méthodes évaluent les gènes candidats en utilisant les preuves de co-occurrence avec une certaine maladie identifié 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** [4, 184, 29, 108, 54, 206, 203, 94]. C'est aujourd'hui la méthode la plus répandue dans la communauté de priorisation des gènes candidats et compte la célèbre méthode Endeavour [4]. 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 BLAST. 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** [202, 103, 107, 100, 101, 86, 165, 144]. 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 complétion de matrice dans les systèmes de recommandation [127, 203]. 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

7. https://www.w3.org/Submission/SWRL/
8. http://spinrdf.org/spin.html
9. https://www.w3.org/TR/shacl

ce jour [203].

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 clefs 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 [203]. 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 [126]. 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 inclurai 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 trouvé 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.

5.4.2 Analyse fonctionnelle des réseaux d'interaction moléculaires

Le succès récent des modèles de graphes et de l'apprentissage profond (deep learning) en bioinformatique [205, 109, 41, 88, 201] suggère la possibilité d'incorporer systématiquement de multiples sources d'information dans le réseau hétérogène et d'apprendre la relation non linéaire entre les phénotype 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.

Parce que RDF est un modèle de représentation basé sur un modèle multi-graphe orienté étiqueté, AgroLD contiendra plusieurs graphes représentant la connaissance par domaine. Par exemple, les signaux de transduction dans les voix métaboliques, les réseaux de régulation de gènes ou les réseaux d'interaction protéine-protéine. Toutefois, il existe peu d'outils et d'algorithmes capables de gérer l'analyse d'énormes réseaux.

Récemment, de nouvelles approches combinant graphes de connaissances et apprentissage profond, ont été proposées dans le domaine du Web sémantique [149, 31, 93, 132]. Je compte approfondir ces aspects afin de proposer une méthode qui inclurait les graphes de connaissances d'AgroLD.

5.5 Conclusion

A travers ce projet de recherche, j'ai essayé de décrire le développement de méthodes permettant de gérer et structurer la complexité des données biologiques afin d'en extraire de la connaissance. Cette connaissance enrichie ayant pour objectif d'identifier les mécanismes moléculaires

contrôlant l'expression de phénotypes chez les plantes. La finalité de ce projet même s'il est de nature pluri-disciplinaire reste avant tout tourné vers l'amélioration des connaissances biologiques.

L'objectif de ce projet sera en effet, de développer des approches de priorisation de gènes candidats utilisant un réseau d'interaction 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. 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 seront directement dédiés aux scientifiques de l'unité Diade mais également pour ses collaborateurs, car il existe actuellement un réel verrou dans la gestion et le traitement des données biologiques.

Ce projet devrait 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) et l'AfricaRice dans le cadre du workpackage « Big Data integration platform » du projet international Rice CRP. Il trouve également une place dans l'initiative Européenne Elixir-Exelerate sur le développement d'une infrastructure Bioinformatique pour l'échange de données et de services sur le Web.

Bibliographie

[1] 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.

[2] 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.

[3] 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.

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

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

[6] 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.

[7] 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.

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

[9] 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.

[10] Joachim BARAN et al. « GFVO : the Genomic Feature and Variation Ontology. » In : *PeerJ* 3 (2015), e933. ISSN : 2167-8359. DOI : 10.7717/peerj.933.

[11] 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.

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

[13] 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.

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

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

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

[17] 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).

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

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

[20] 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.

[21] 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).

[22] 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.

[23] 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.

[24] 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.

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

[26] 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.

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

[28] 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.

[29] 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.

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

[31] 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.

[32] Sarah COHEN-BOULAKIA et Ulf LESER. « Next generation data integration for Life Sciences ». In : (2011), p. 1366–1369.

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

[34] 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.

[35] 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.

[36] 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.

[37] 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.

[38] 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.

[39] 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.

[40] Richard (National University of Ireland) CYGANIAK et Christian BIZER. « Pubby - A Linked Data Frontend for SPARQL Endpoints ». In : (2008). Citation Key : Cyganiak2008.

[41] 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.

[42] 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.

[43] 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.

[44] K. H. DAVIS et A. K. ARORA. « Converting A Relational Database Model into an Entity-Relationship Model ». In : *in Proceedings of the Sixth International Conference on Entity-Relationship Approach*. 1987.

[45] 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.

[46] 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.

[47] 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.

[48] 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.

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

[50] 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.

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

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

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

[54] 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).

[55] 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.

[56] 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).

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

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

[59] Chiara GABELLA, Christine DURINX et Ron APPEL. « Funding knowledgebases : Towards a sustainable funding model for the UniProt use case ». en. In : *F1000Research* 6 (nov. 2017), p. 2051. ISSN : 2046-1402. DOI : 10.12688/f1000research.12989.1. (Visité le 24/03/2019).

[60] 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.

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

[62] 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.

[63] 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.

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

[65] 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.

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

[67] 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.

[68] 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.

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

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

[71] 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.

[72] 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.

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

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

[75] 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.

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

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

[78] 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.

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

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

[81] 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.

[82] 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.

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

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

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

[86] 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.

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

[88] 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.

[89] 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.

[90] 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. DOI : 10.1093/jxb/erv271.

[91] 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.

[92] 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.

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

[94] 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.

[95] 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).

[96] 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.

[97] 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.

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

[99] 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.

[100] 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.

[101] 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.

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

[103] 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.

[104] 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.

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

[106] « 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.

[107] 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.

[108] 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.

[109] 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.

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

[111] 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.

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

[113] 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.

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

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

[116] 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).

[117] 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).

[118] 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.

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

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

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

[122] 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.

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

[124] 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.

[125] 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.

[126] 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.

[127] 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.

[128] Theodor H NELSON. « From Memex to Hypertext ». In : 1991. Chap. As We Will Think, p. 245–260.

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

[130] 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.

[131] 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.

[132] 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.

[133] 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.

[134] 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.

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

[136] 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.

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

[138] 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.

[139] Paul OTLET. « Traité de documentation ». In : Mundaneum, 1934, p. 40,205,377–381,428.

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

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

[142] 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.

[143] 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.

[144] 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.

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

[146] Leonore REISER et al. « Sustainable funding for biocuration : The Arabidopsis Information Resource (TAIR) as a case study of a subscription-based funding model ». eng. In : *Database : The Journal of Biological Databases and Curation* 2016 (2016). ISSN : 1758-0463. DOI : 10.1093/database/baw018.

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

[148] 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.

[149] 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.

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

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

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

[153] 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.

[154] 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).

[155] Max SCHMACHTENBERG, Christian BIZER et Heiko PAULHEIM. « Adoption of the Linked Data Best Practices in Different Topical Domains ». In : *Semantic Web Conference* 1 (2014), p. 245–260.

[156] 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.

[157] 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.

[158] 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.

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

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

[161] 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.

[162] 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.

[163] 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.

[164] 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).

[165] 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.

[166] 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).

[167] 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).

[168] 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.

[169] 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.

[170] 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.

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

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

[173] 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.

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

[175] 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.

[176] 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.

[177] 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.

[178] 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.

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

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

[181] 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.

[182] 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.

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

[184] 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).

[185] 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.

[186] 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.

[187] 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.

[188] 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.

[189] Pierre-Yves VANDENBUSSCHE. « Thèse : Définition d'un cadre formel de représentation des Systèmes d'Organisation de la Connaissance ». In : 2011, p. 14–19,58–61.

[190] 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.

[191] 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).

[192] Patricia L WHETZEL et al. « The MGED ontology : A Resource for semantics-based description of microarray experiments ». In : 22.7 (2006), p. 866–873.

[193] Gio WIEDERHOLD et Michael R. GENESERETH. « The basis for mediation ». In : *Proc. COOPIS* (1995). (Visité le 02/02/2013).

[194] Gio WIEDERHOLD et Michael R. GENESERETH. « The Conceptual Basis for Mediation Services ». In : *IEEE Expert* 12.5 (1997), p. 38–47.

[195] 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.

[196] Mark D WILKINSON et Matthew LINKS. « BioMOBY : an open source biological web services proposal. » In : *Briefings in Bioinformatics* 3.4 (2002), p. 331–341.

[197] Mark D. WILKINSON et al. « The FAIR Guiding Principles for scientific data management and stewardship ». In : *Scientific Data* 3 (2016), p. 160018.

[198] Antony J. WILLIAMS et al. *Open PHACTS : Semantic interoperability for drug discovery*. 2012.

[199] 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.

[200] 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.

[201] 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).

[202] 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).

[203] 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.

[204] 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).

[205] 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.

[206] 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

Liste des tableaux

Annexe A

Article 1

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

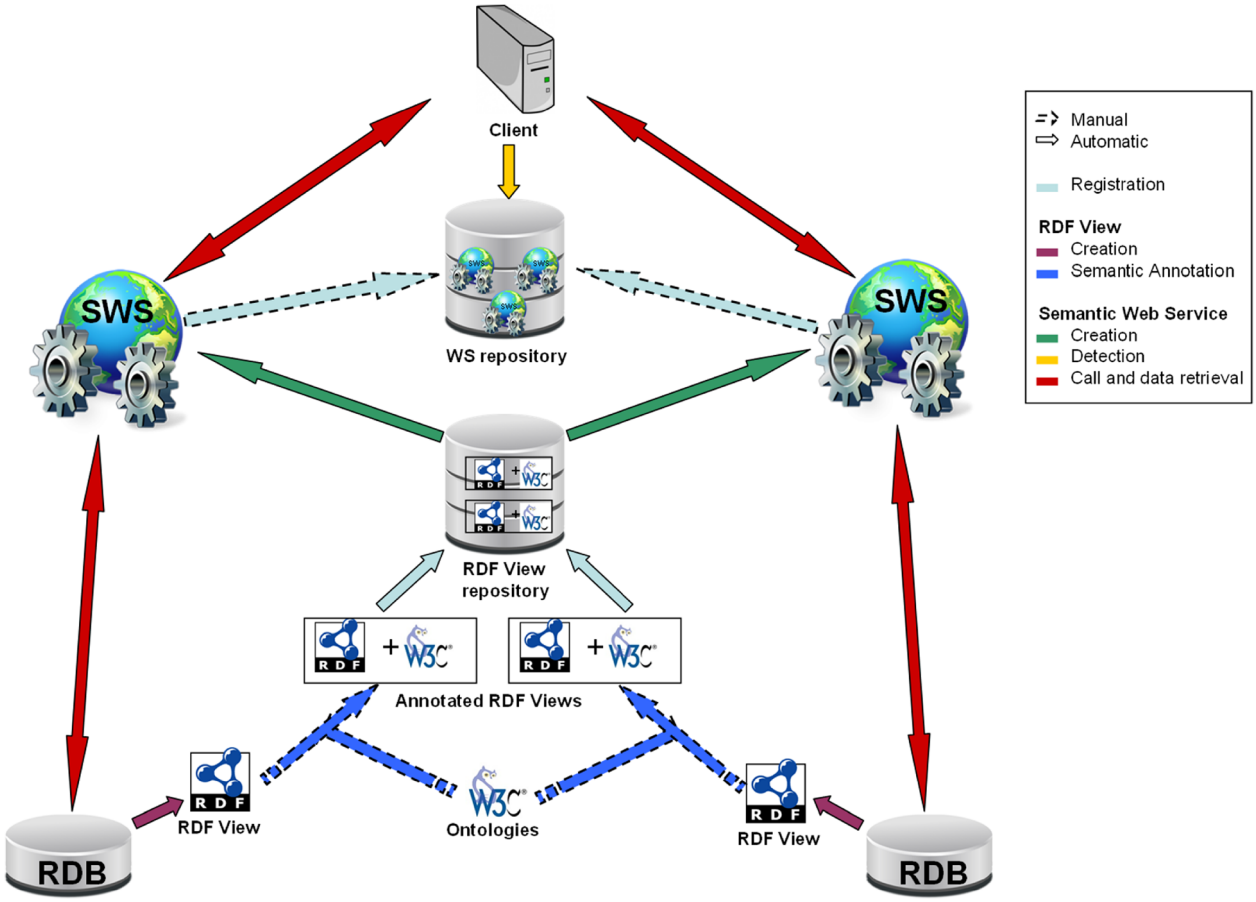


Figure 1 Global architecture of the BioSemantic framework. The first contribution of our work is the automatic creation of an RDF view containing RDF metadata, which is necessary for the automatic creation of Semantic Web Services. The second contribution is the automatic creation and deployment of Semantic Web Services.

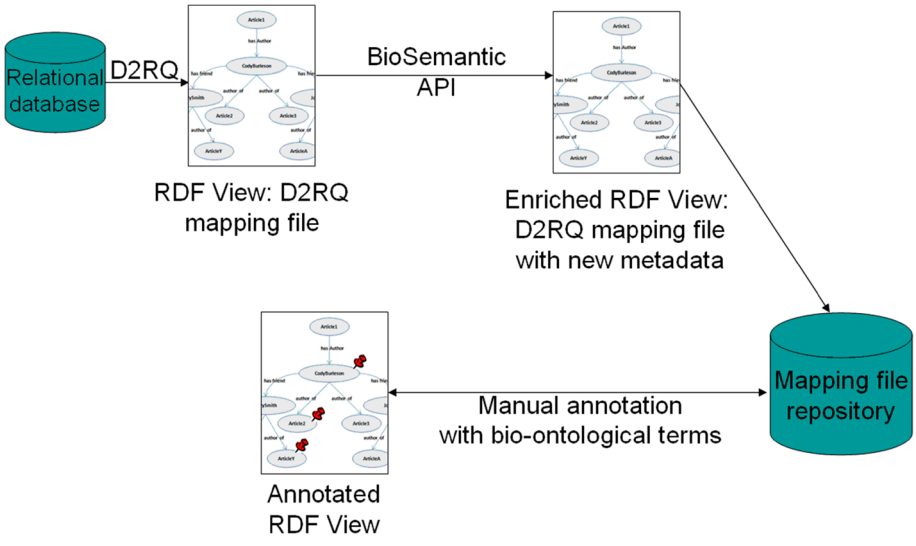


Figure 2 Generation and semi-automatic annotation of the RDF view. D2RQ creates the D2RQ mapping file, and our BioSemantic API automatically adds new metadata about the database schema. Finally, the mapping file is stored in a repository. Annotation with bio-ontological terms is performed manually by an expert.

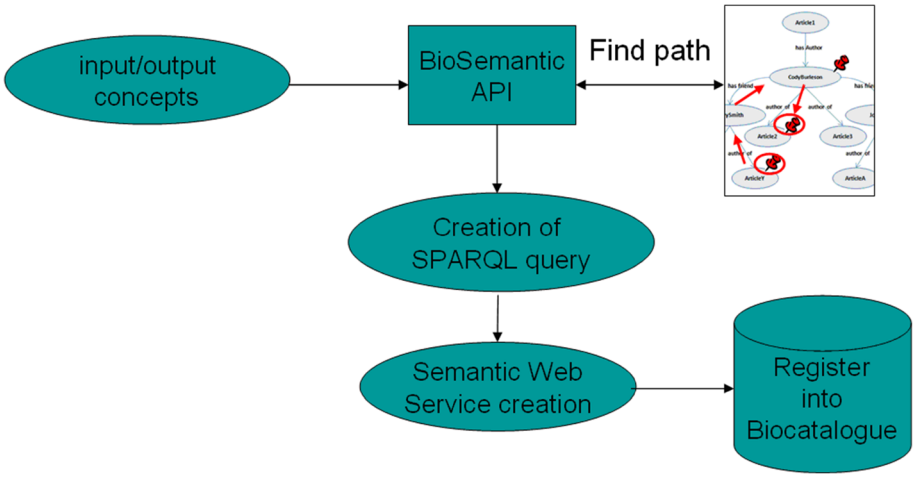


Figure 3 Automatic generation of Semantic Web Services. The Web Services developer selects the bio-ontological terms to be used as input/output. All of the mapping files, which are stored in the mapping file repository, are automatically parsed to find a path linking the input and output ontological 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

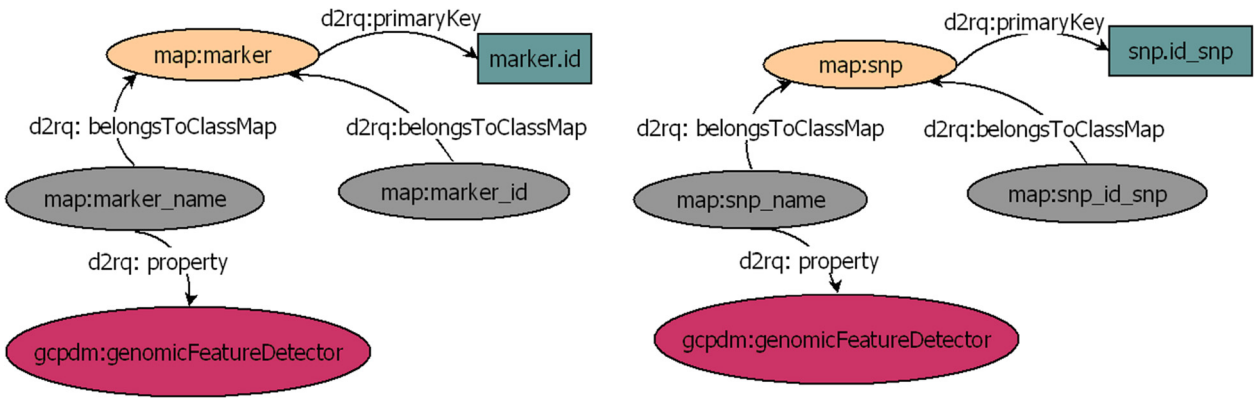


Figure 4 Graph-based representation of annotated RDF views. Each graph is the RDF representation of some part of a relational database. The *d2rq:belongsToClassMap* property links a column to a table. The *d2rq:primaryKey* property defines the primary key of a table. The *d2rq:property* property links 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

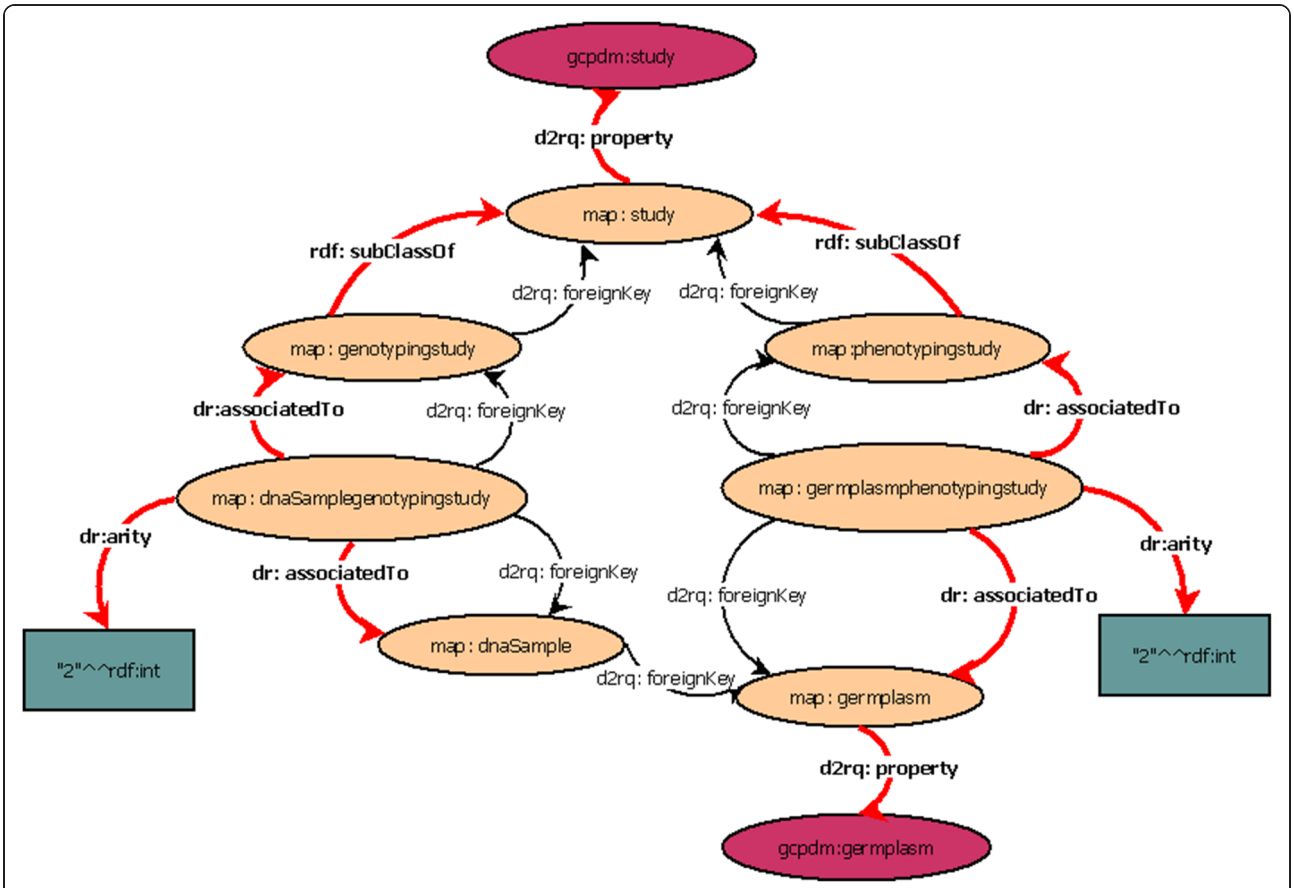


Figure 5 Classification of the database table relationships. Each light node represents a table of the relational database. Here, we only show the tables, and the columns are not represented. The dark nodes represent the semantic annotations. Each edge represents a property that is shared between 2 nodes. 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.

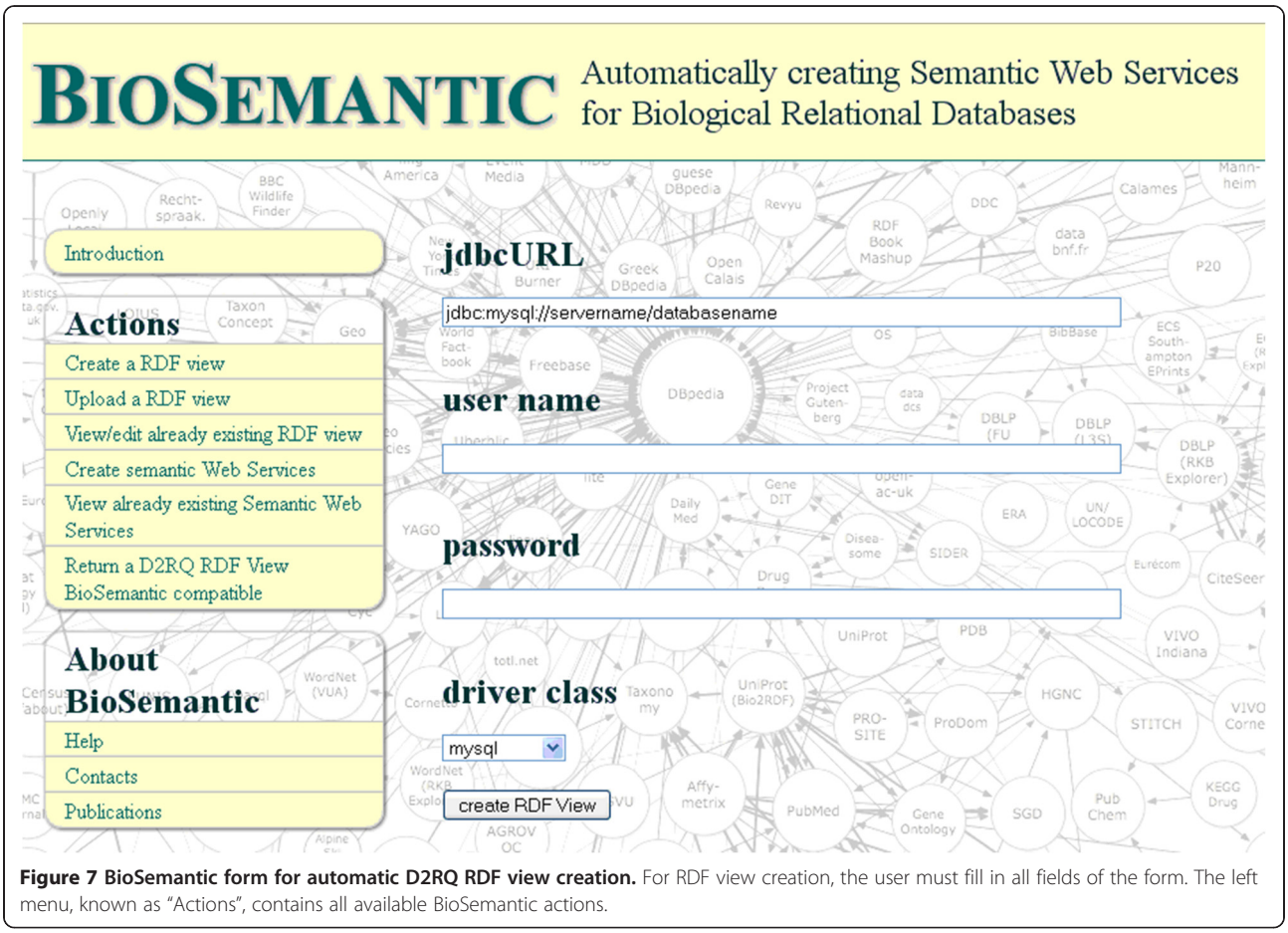


Figure 7 BioSemantic form for automatic D2RQ RDF view creation. For RDF view creation, the user must fill in all fields of the form. The left menu, known as "Actions", contains all available BioSemantic actions.

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

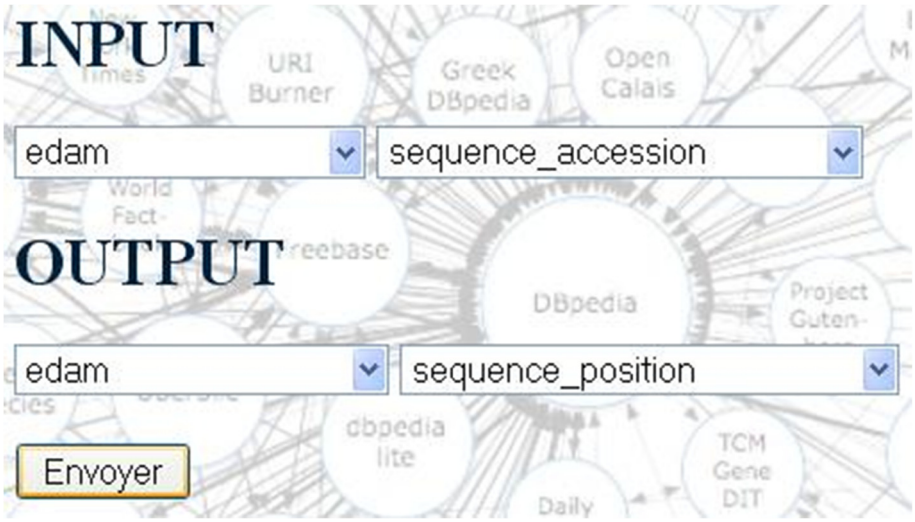


Figure 8 BioSemantic form for input/output concept selection. These concepts will be used to detect a path and to annotate the input/output of the SWS. The user can only select the prefix and concepts used to annotate a previously registered RDF view.

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.

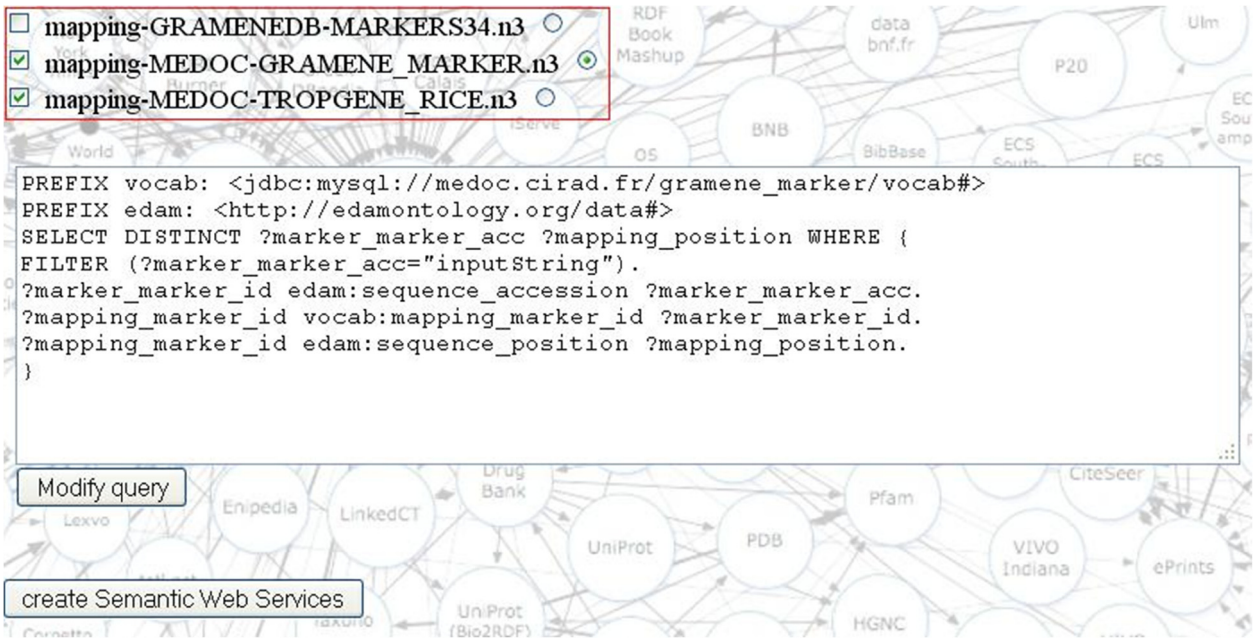


Figure 9 BioSemantic form for RDF view selection and query visualisation/edition. The red rectangle contains the name of the RDF views annotated with both input/output concepts. The checkbox before each name allows for the selection of a view for SWS creation. The radio button after the name of an RDF view allows for query visualisation/edition. When all desired RDF views are selected, a simple click creates the SWS.

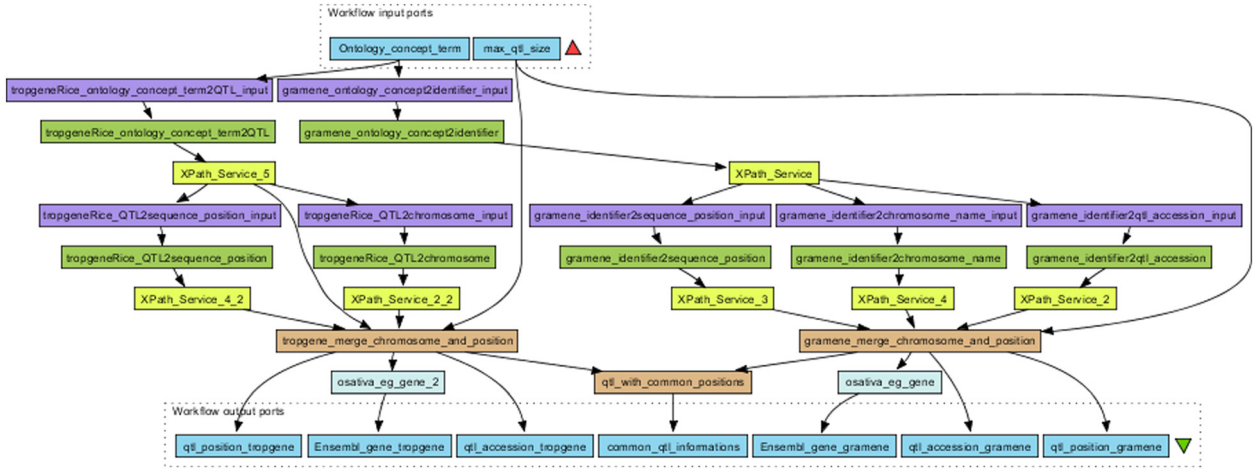


Figure 10 Use case workflow created with Taverna 2. The workflow contains BioSemantic SWS querying for both the TropGene and Gramene databases for QTL information retrieval and also contains BioMart WS querying EnSembl for gene information retrieval. This workflow also detects QTLs from TropGene and 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.

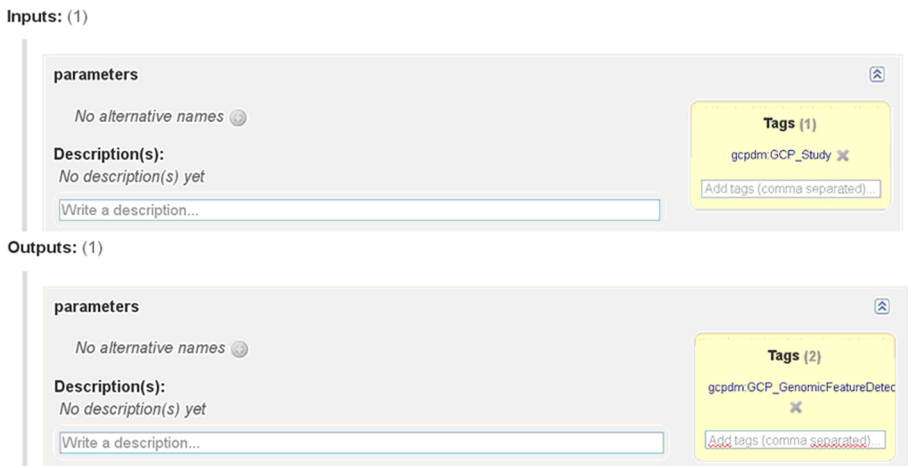


Figure 12 BioCatalogue input/output annotations for the GetTropgeneMarkerSPARQL semantic Web Service. Our bio-ontological terms correspond to the input/output tags in the BioCatalogue.

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.

Annexe B

Article 2

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

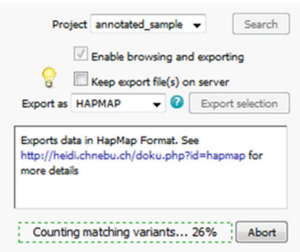


Fig. 1 Action panel enabling project selection, progress indication, abort and export functionalities

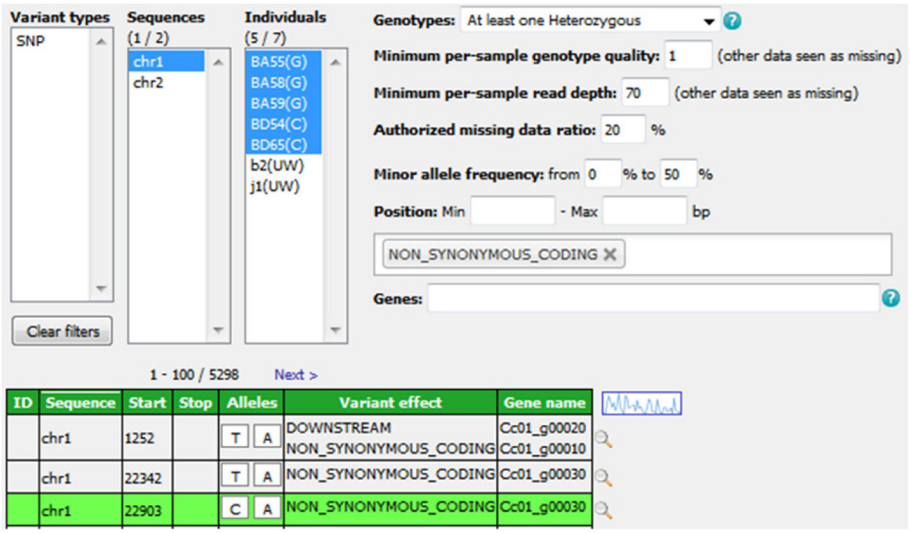


Fig. 2 Variant-filtering interface, allowing search criteria definition and result browsing

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;

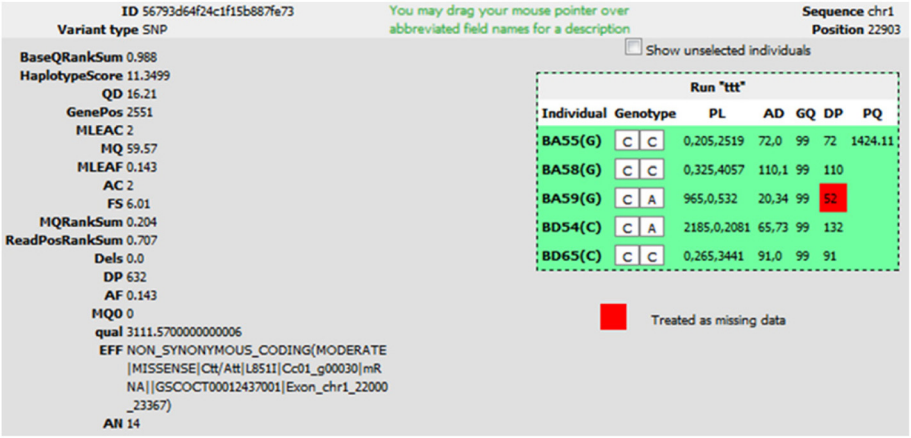


Fig 3 Variant detail dialogue, providing variant metadata and genotype-level information

- 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

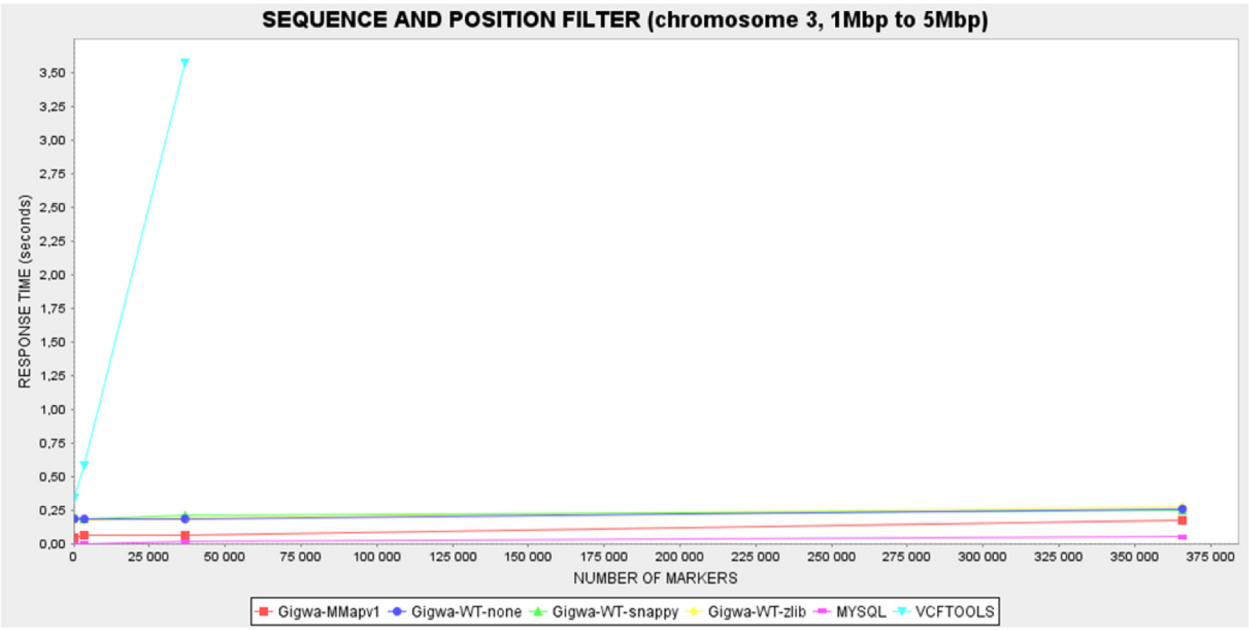


Fig. 4 Response-time plot by tool for first benchmarked query (location-based filter). VCFtools is by far the slowest option

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

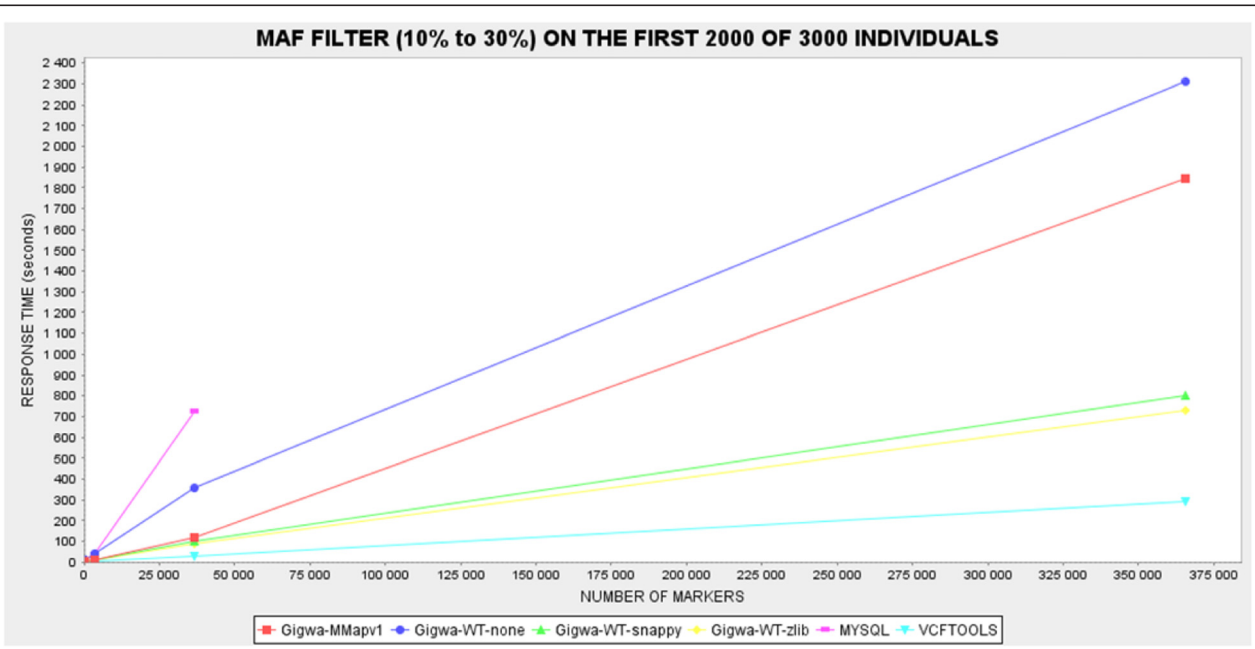


Fig. 5 Response-time plot by tool for second benchmarked query (MAF filter). MySQL is by far the slowest option

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

Annexe C

Article 3

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

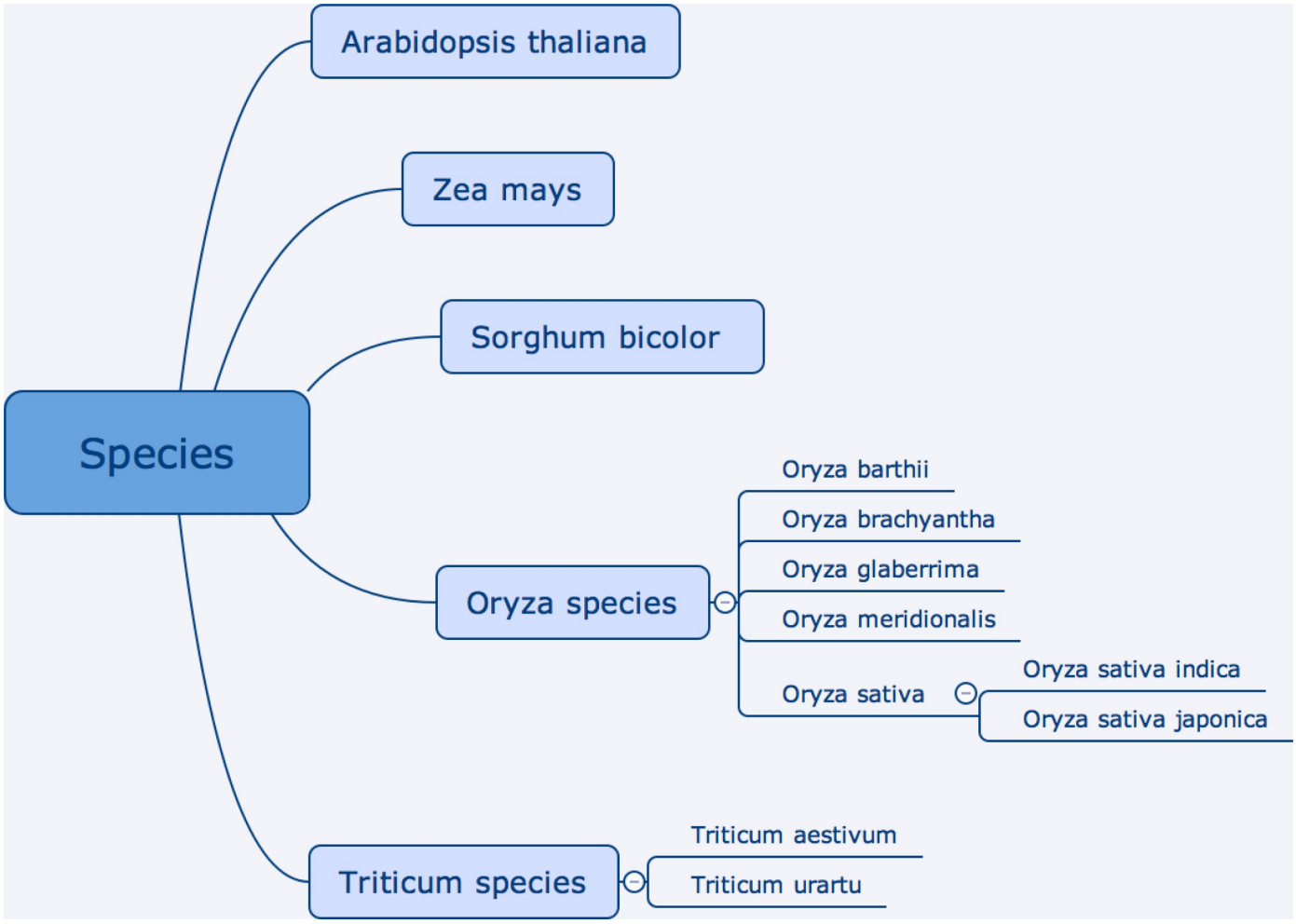


Fig 1. Current plant species included in AgroLD.

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

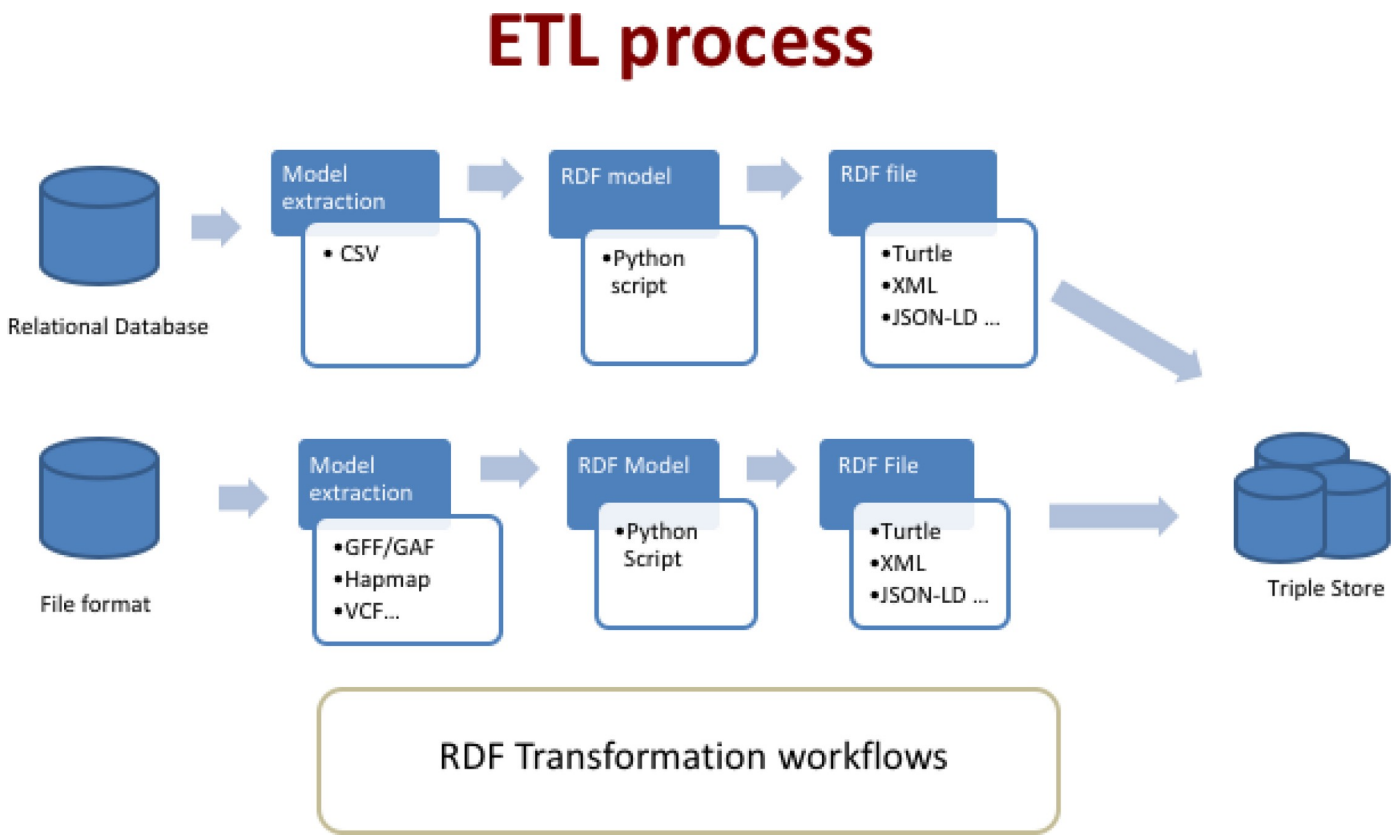


Fig 2. ETL workflow for the various datasets and data formats. The workflow shows two types of process: 1) from relational databases through a CVS file export: in that case, the transformation is tailored for the database model with some Python scripts converters. 2) from standards file formats: in that 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

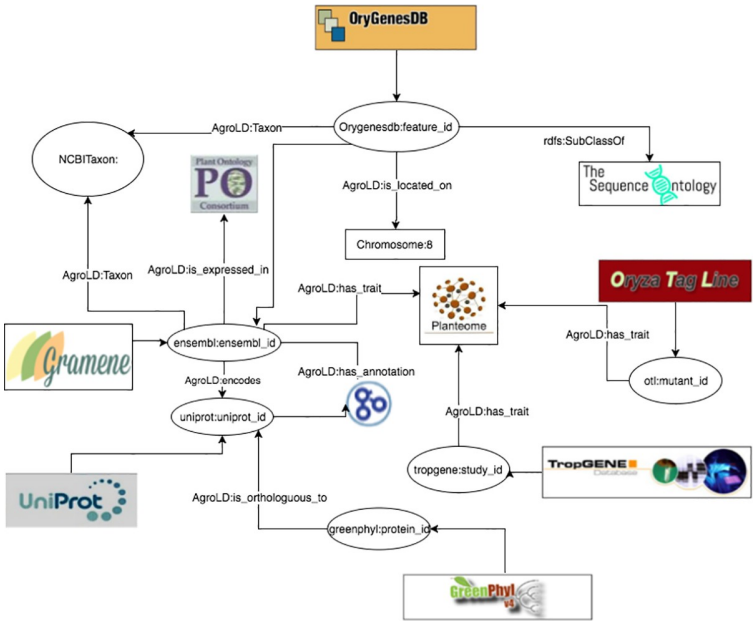


Fig 3. Linking information in AgroLD. The figure illustrates the linking of varies information in AgroLD.

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

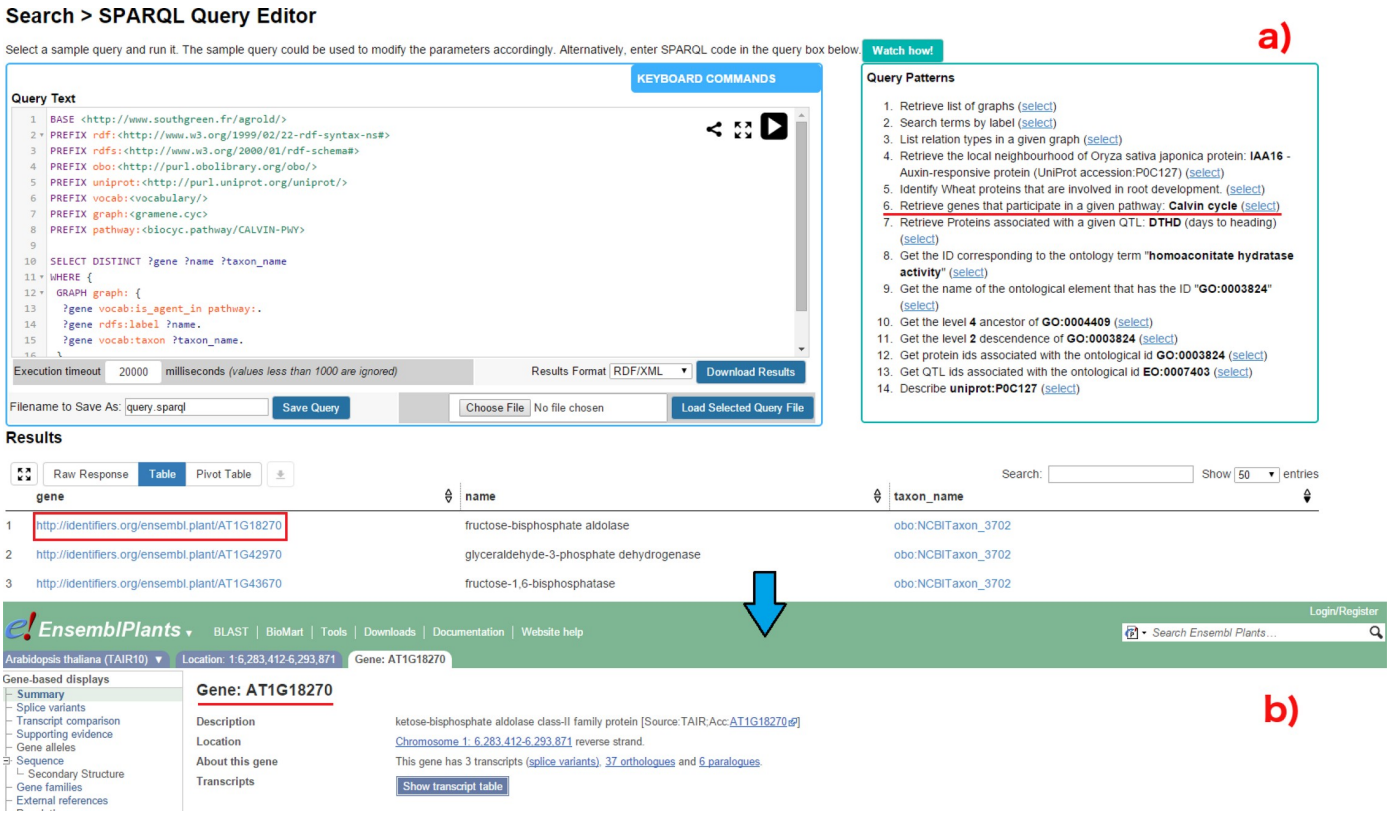


Fig 4. SPARQL query editor. Figure illustrates the execution of query Q6: (a) Q6 is one the examples queries on the top-right corner (highlighted in red). On executing the query, the results are rendered below the editor; (b) the user can look up specific genes of interest by clicking on the corresponding 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:

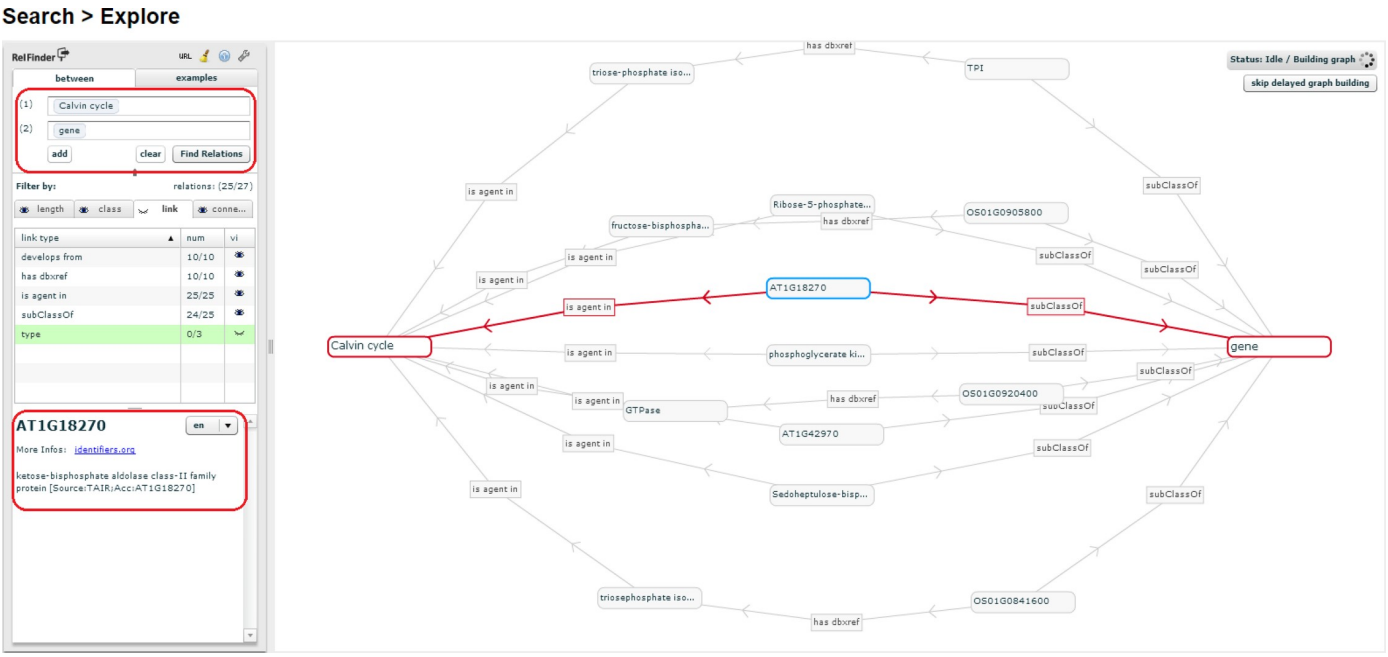


Fig 5. Exploring entity relationships in AgroLD. Figure illustrates differently the results obtained for Q6 using Explore Relationships tool. The results of Q6 can be visualized by entering the concepts (Calvin cycle and gene) in the left panel. On executing the query, all the genes involved in the chosen 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*

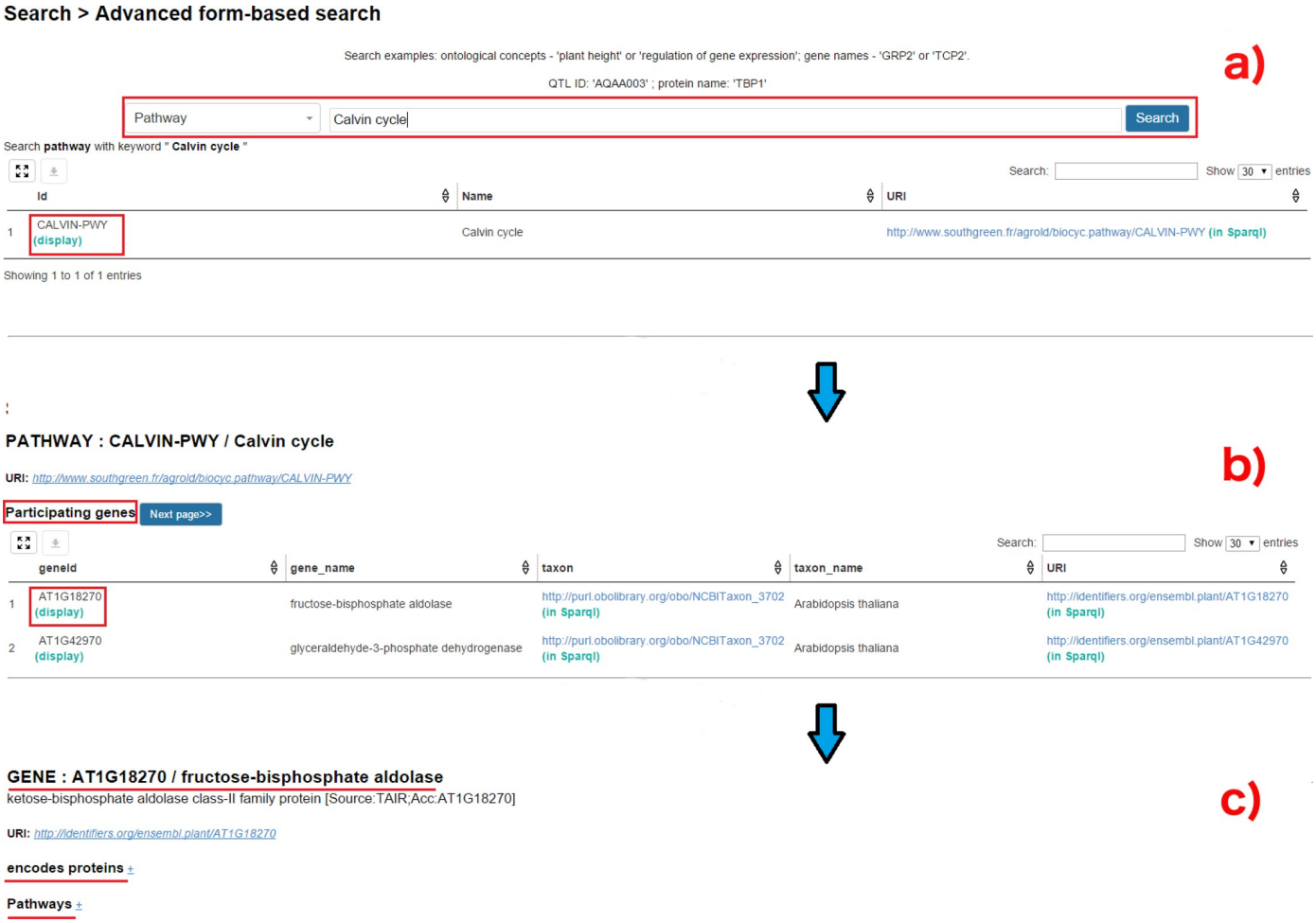


Fig 6. Advanced search query form: Figure demonstrates the steps involved in retrieving the results for Q6 using the Advanced Search query form: (a) query Q6 can be executed by selecting the type of entity (Pathways–highlighted in red) to search and entering the name of the entity (Calvin cycle). The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