Nadia Yacoubi Ayadi, Catherine Faron, Franck Michel, Robert Bossy, Arnaud Barbe

Construction et exploitation d'un graphe de connaissances sur la littérature scientifique en sciences de la vie

Volume 6, nº 1-2 (2025), p. 107-129.

https://doi.org/10.5802/roia.95

La Revue Ouverte d'Intelligence Artificielle est membre du Centre Mersenne pour l'édition scientifique ouverte
www.centre-mersenne.org
e-ISSN : 2967-9672

Revue Ouverte d'Intelligence Artificielle
Volume 6, nº 1-2, 2025, 107-129

Construction et exploitation d'un graphe de connaissances sur la littérature scientifique en sciences de la vie

Nadia Yacoubi Ayadi[a], Catherine Faron[b], Franck Michel[b], Robert Bossy[c], Arnaud Barbe[b]

[a] Université de Lyon 1, CNRS, LIRIS, UMR 5205 (France)
E-mail : nadia.yacoubi-ayadi@univ-lyon1.fr

[b] Université Côte d'Azur, Inria, CNRS, I3S, UMR 7271 (France)
E-mail : faron@i3s.unice.fr, fmichel@i3s.unice.fr

[c] MaIAGE, INRAE, Université Paris-Saclay, 78350 Jouy-en-Josas (France)
E-mail : robert.bossy@inrae.fr, arnaud.barbe@i3s.unice.fr.

Résumé. — Dans cet article, nous présentons un graphe de connaissances construit à partir d'un corpus d'articles scientifiques portant sur les méthodes de sélection génomique pour la culture du blé. Ces méthodes contribuent à l'amélioration du profil agronomique et de la qualité des variétés de blé. La littérature scientifique sur le sujet est en croissance continue ces vingt dernières années. Dans un premier temps, un outil de traitement automatique du langage nous a permis d'extraire et de normaliser différentes entités nommées en les associant à des concepts préalablement définis dans une ontologie du domaine. Ces entités se réfèrent à des noms de gènes, traits, phénotypes, marqueurs et variétés (cultivars) de blé. Nous avons construit un graphe qui intègre, structure et décrit ces entités en se basant sur l'ontologie W3C Web Annotation Ontology (OA) pour formaliser la description du contexte d'apparition de ces entités au sein du corpus et apporter ainsi des indications sur les liens et les associations fréquentes entre ces éléments. En s'appuyant sur un ensemble de questions de compétence formulées par un expert du domaine, nous avons validé la pertinence du modèle proposé et par conséquent le graphe de connaissances généré. Afin de rendre notre graphe accessible à un grand nombre d'utilisateurs, nous avons développé plusieurs interfaces de recherche et de visualisation permettant d'explorer les contextes d'apparition de plusieurs entités (de différents types) dans le même article. Ce travail contribue à la structuration, à la compréhension et l'exploration des connaissances dans le domaine de la sélection génomique du blé, en fournissant un cadre formel pour la découverte et l'analyse des relations entre les entités pertinentes du domaine. Nous proposons une méthode d'ingénierie des connaissances de bout en bout, générique et adaptable à la valorisation d'un corpus de littérature scientifique quel que soit son domaine scientifique.

Mots-clés. — Données liées, ontologies, annotation sémantique, graphes de connaissances, fouille de textes.

1. Introduction

La culture du blé est l'une des plus importantes et répandues dans le monde, elle fournit le principal apport de protéines pour une grande part de la population mondiale. Les sélectionneurs cherchent à obtenir des variétés aux propriétés intéressantes pour la productivité, la résistance aux maladies et l'adaptation aux changements climatiques. Les techniques modernes de phénotypage et de dépistage génomique permettent une sélection ciblée et une meilleure hybridation des variétés de blé. Ces techniques rendent possible l'obtention de graines résistantes aux maladies et à la sécheresse, tout en étant productives et durables. Elles combinent une recherche génétique fondamentale en laboratoire et une expérimentation sur terrain. Une partie des résultats de ces recherches est enregistrée dans des bases de données génomiques libres d'accès. En revanche, une autre partie n'est accessible qu'à travers l'exploration de la littérature scientifique. Cependant, il est impossible pour un chercheur de parcourir l'ensemble des publications scientifiques vu leur volume exponentiel. Par conséquent, les techniques de Traitement Automatique de Langues (TAL) ont été largement utilisées pour la fouille de textes dans l'objectif d'extraire les informations pertinentes pouvant aider les chercheurs dans leurs investigations.

Dans le contexte des sciences de la vie, ces outils de TAL permettent de reconnaître automatiquement les entités pertinentes du domaine. De façon complémentaire/orthogonale, les technologies du Web sémantique offrent un cadre avantageux pour gérer, structurer, agréger et connecter un ensemble d'entités hétérogènes dans un graphe de connaissances (KG) [9]. Les entités qu'elles soient des entités génomiques ou publications scientifiques sont identifiables par des URI. Leurs relations, comme les interactions entre les gènes, les mentions des traits agronomique dans la littérature, ou la description des thèmes des articles par des descripteurs, se décrivent en RDF.

Ce travail de recherche a été mené dans le cadre du projet D2KAB[(1)], un projet de recherche ANR ayant pour objectif de créer un cadre pour transformer les données d'agronomie et de biodiversité en connaissances décrites sémantiquement, interopérables, exploitables et ouvertes. L'objectif de notre travail s'aligne avec l'objectif du projet D2KAB et vise la construction d'un graphe de connaissances RDF à partir d'un corpus scientifique. En première phase, l'outil AlvisNLP[(2)] développé par l'équipe MaIAGE nous a permis d'extraire et d'annoter (lier) les différents types d'Entités Nommées (EN) en utilisant des concepts préalablement définis dans des ontologies du domaine. Ceci permet d'une part de désambiguïser les entités nommées extraites et d'autre part de tirer profit de la structure des ontologies domaine. Le graphe résultat que nous nommons WheatGenomics-SLKG pour la génomique de blé répond au besoin de structurer et d'intégrer les entités extraites à partir du texte, tout en adoptant un modèle adéquat pour représenter leurs contextes d'apparition dans le texte. Ceci permet de répondre à des questions complexes telles que l'identification de marqueurs génétiques liés à des gènes impliqués dans le contrôle d'un trait agronomique d'intérêt.

[(1)] http://www.d2kab.org/

[(2)] https://bibliome.github.io/alvisnlp/

Guidés par un ensemble de questions de compétences (CQ), nous avons proposé un modèle qui réutilise des ontologies et des vocabulaires existants pour structurer et représenter de façon uniforme à la fois les publications scientifiques et leurs méta-données et les annotations d'entités nommées dans le même graphe de connaissances. Le processus de construction du graphe est réalisé en deux phases parallèles : (1) l'utilisation de l'outil Morph-xR2RML [21, 27] pour transformer les annotations d'EN produites par l'outil AlvisNLP en un jeu de données RDF que nous avons publié via un point d'accès SPARQL public[(3)] (2) l'enrichissement du graphe avec des méta-données descriptives des publications scientifiques à partir de l'API PMC Entrez[(4)] et de les intégrer dans le graphe. Enfin, toutes les CQ ont été traduites en requêtes SPARQL et évaluées pour valider le graphe de connaissances obtenu et le modèle sémantique sous-jacent ; les résultats des requêtes ont été validés par les experts. De plus, nous avons développé deux interfaces de visualisation permettant à des utilisateurs non experts des technologies du Web sémantique d'explorer et d'interroger le graphe obtenu. La première utilise les entités nommées comme filtres de recherche d'articles et permet de visualiser toutes les entités dans le résumé d'un article sélectionné. La seconde interface est destinée à offrir à l'utilisateur une approche visuelle adaptée à la résolution de questions de compétences complexes. Cet article est une extension de notre publication dans les actes de la conférence IC 2022 [38] dans laquelle nous présentions une première version du graphe de connaissances WheatGenomics-SLKG. Dans le présent article nous présentons une version consolidée de ce graphe, maintenant accessible à travers un SPARQL end-point public[(5)], accompagné d'un ensemble enrichi de questions de compétence implémentées en SPARQL et disponibles sur le github du projet[(6)] et d'interfaces de visualisation que nous avons développées pour son exploration et interrogation.

Notre travail répond aux enjeux de la valorisation des corpus de littérature scientifique dont le volume va croissant et nécessite le développement de méthodes et outils d'ingénierie des connaissances pour leur exploitation. Nous proposons une méthode générique et adaptable à tout domaine scientifique. Il s'agit d'une méthode d'ingénierie des connaissances de bout en bout, allant de l'extraction des connaissances à partir de textes à l'aide d'outils de traitement automatique de la langue, jusqu'à l'exploration et l'analyse visuelles du corpus des publications scientifiques, en requêtant le graphe de connaissances capturant les annotations sémantiques des textes produites, en passant par la définition d'un modèle générique de représentation de ces annotations. L'implémentation de cette méthode repose sur la combinaison d'un ensemble d'outils au sein d'une chaîne de traitements unifiée. Cette approche permet notamment d'exploiter les cooccurrences entre entités afin d'identifier et de révéler des réseaux d'interactions potentielles. En facilitant l'exploration et l'analyse de grands corpus de littérature scientifique, cette méthodologie favorise l'interprétation riche et contextualisée des

(3) `http://d2kab.i3s.unice.fr/sparql`
(4) `https://www.ncbi.nlm.nih.gov/pmc/tools/developers/`
(5) `http://d2kab.i3s.unice.fr/sparql`
(6) `https://github.com/Wimmics/WheatGenomicsSLKG`

contenus scientifiques et la génération d'hypothèses scientifiques, et contribue à la découverte de nouvelles connaissances.

Cet article est organisé comme suit. Dans la section 2, nous présentons une synthèse (non exhaustive) des approches de construction de graphes de connaissances à partir de textes scientifiques ; ainsi que les vocabulaires réutilisés dans ce travail de recherche. Dans la section 3, nous présentons un ensemble de CQ qui résument les exigences et les besoins potentiels des experts d'exploiter les annotations générées. Le modèle sémantique du graphe de connaissances est présenté dans la section 4. Dans la section 5, nous détaillons le processus et les outils que nous avons utilisés pour la génération du graphe de connaissances. Dans la section 6, nous présentons des requêtes SPARQL qui correspondent à l'implémentation de certains CQ présentées dans la section 3. Dans la section 7 nous présentons les interfaces de visualisation développées.

2. État de l'art

Dans cette section, nous discutons quelques approches existantes pour la construction de graphes de connaissances à partir de textes. Ensuite, nous présentons les vocabulaires et les ontologies que nous avons réutilisés pour structurer les connaissances dans le futur graphe, à savoir : les ontologies FaBio (the FRBR-aligned Bibliographic Ontology) [31], BIBO (BIBliographic Ontology) et le vocabulaire Web Annotation Vocabulary (OA) [34].

2.1. Construction de graphes de connaissances à partir de textes

Ces dernières années, plusieurs graphes de connaissances ont été construits pour répondre aux besoins d'exploration de la littérature scientifique et d'intégration de données dans le domaine des sciences de la vie [4, 15, 26]. Le travail de recherche présenté dans [13] décrivant l'une des tâches du challenge BioNLP-ST 2013 a été l'un des premiers à mettre en évidence l'intérêt de construire des graphes de connaissances RDF à partir d'annotations extraites à partir de textes. Pour ce challenge, les bases de connaissances RDF construites à partir des annotations extraites par 10 systèmes TAL ont été évaluées. Les approches de construction de graphes à partir de données textuelles reposent sur des chaînes de traitements implémentant des tâches clés, souvent réalisées de manière semi-automatique et/ou incrémentale [8]. Ces tâches incluent notamment la reconnaissance et le liage d'entités nommées (REN et LEN) ainsi que l'extraction de relations (ER), structurant les entités dans le graphe. Pour les tâches d'extraction et de liage d'entités nommées, plusieurs approches ont été proposées, les premières sont celles basées sur des dictionnaires, comme les lexiques ou les gazetteers et qui nécessitent des ressources de haute qualité pour maximiser leur couverture [1, 35]. Comme pour de nombreuses tâches de TAL, les méthodes basées sur l'apprentissage automatique et profond deviennent prédominantes [7, 30]. Ces méthodes peuvent être combinées ensemble pour améliorer la qualité des résultats. Actuellement, ces méthodes sont en passe d'être dépassées par l'utilisation des grands modèles de langue (LLM) pré-entraînés sur une très grande quantité de texte [18]. L'objectif de l'extraction de relations est de découvrir les relations d'intérêt entre une paire d'entités,

décrivant ainsi leur interaction [10]. Pour la tâche d'extraction de relations à partir de textes les approches proposées sont basées sur des règles [33] ou sur l'apprentissage automatique (supervisées [17] et non supervisées [6]). Plus récemment, des approches de bout en bout (*End-to-End Relation Extraction*) ont été proposées pour s'attaquer simultanément aux tâches de REN et ER. Dans ce scénario, un modèle est entraîné simultanément sur les objectifs de REN et de ER [11]. Malgré ces avancées, plusieurs défis subsistent. L'une des principales limitations réside dans la disponibilité réduite de jeux de données annotés manuellement par des experts, en particulier pour certaines catégories d'entités nommées telles que les traits et les phénotypes des plantes. Ceci entrave l'évaluation des tâches de REN, LEN et ER et par conséquent la qualité des graphes de connaissances produits à partir de textes. Très récemment, le corpus *Taec* a été publié comme le premier jeu de données « *gold standard* » dédié aux traits et aux phénotypes du blé [29]. En ce qui concerne, le liage d'entités, les vocabulaires dans ces domaines sont volumineux et complexes [4]. On observe un décalage important entre les étiquettes de concepts et les mentions dans les textes scientifiques avec notamment l'usage fréquent d'abréviations, de métonymies et de variations syntaxiques ([16], [3]). Enfin, l'évolution rapide et continue des connaissances dans les sciences de la vie impose des mises à jour régulières des bases de connaissances et des modèles d'extraction. Cette dynamique accroît la complexité des efforts de standardisation et d'intégration des données, et le développement d'approches robustes et évolutives devient de ce fait crucial.

2.2. Vocabulaires et Ontologies existants

Pour représenter à la fois les publications scientifiques et les annotations extraites à partir de ces publications, nous avons réutilisé plusieurs vocabulaires et ontologies. D'une part, nous avons adopté les ontologies FaBio (the FRBR-aligned Bibliographic Ontology) [31] et BIBO[(7)] pour représenter les méta-données descriptives et bibliographiques des publications scientifiques. L'ontologie FaBio est une ontologie dérivée du modèle FRBR [5] qui est un vocabulaire RDF publié par l'IFLA (International Federation of Library Association) pour représenter les notices artistiques et bibliographiques. FRBR définit un ensemble exhaustif de classes permettant de modéliser tout type d'oeuvres et de décrire tout le cycle de vie de l'oeuvre de sa création jusqu'à son adaptation et transformation. D'autre part, L'ontologie BIBO est une ontologie minimale qui décrit essentiellement les agents, les documents et les événements qui conduisent à la production d'une œuvre. Les ontologies BIBO et FaBio sont généralement utilisées de façon complémentaire avec d'autres vocabulaires existants tels que Dublin Core[(8)]ou schema.org[(9)]. Enfin, le vocabulaire Web Annotation Vocabulary (OA) [34] est une recommandation W3C qui propose un ensemble de classes et de propriétés RDF pour représenter de manière uniforme les annotations sur le Web dans un format interopérable, d'où l'intérêt de son utilisation [14].

[(7)] `https://github.com/structureddynamics/Bibliographic-Ontology-BIBO`
[(8)] `https://www.dublincore.org/specifications/dublin-core/dcmi-terms/`
[(9)] `https://schema.org/`

2.3. Visualisation de graphes de connaissances

Explorer un graphe de connaissances RDF nécessite préalablement la définition de requêtes SPARQL permettant d'obtenir des résultats provenant d'un ou de plusieurs endpoints. Cependant, d'une part, la définition des requêtes SPARQL peut se révéler complexe et fastidieuse, nécessitant un haut niveau d'expertise, et d'autres part, les interfaces de visualisation doivent s'adapter à la nature des données RDF (entités géographiques, liens de co-citation, etc.). Ainsi, contrairement aux outils de visualisation classiques, utilisant des jeux de données dont la structure et la nature sont connues à l'avance, la visualisation des données liées nécessite un traitement préalable du jeu de données RDF pour extraire les données appropriées en interrogeant un ou plusieurs endpoints SPARQL. Ces dernières années, un intérêt croissant a été porté au développement d'outils de visualisation permettant aux utilisateurs d'interroger, explorer et interagir avec des jeux de données liées [2]. À travers des techniques de recherche multi-facettes combinant à la fois différents critères et niveaux d'abstraction, ces outils permettent à leurs utilisateurs d'explorer les concepts et les relations pertinents d'un domaine d'application via la représentation d'une ontologie (WebVOWL [19]), d'inspecter des graphes RDF, afin de "*débugguer*" les triplets et d'inférer le schéma ontologique (LD-VOWL [36]), de visualiser les instances sur la base de leurs types/-classes tout en considérant à la fois les dimensions spatiales et temporelles [12, 32, 37]. Dans [20], les auteurs présentent un pipeline de visualisation générique de données liées en implémentant l'outil LDViz comme preuve de concept. Cet outil intègre une interface de gestion des requêtes SPARQL, un moteur de transformation des données et une interface pour permettre la visualisation de données extraites d'un endpoint SPARQL. Ainsi, cet outil permet à tout utilisateur d'accéder au endpoint SPARQL de son choix, d'effectuer des recherches avec des requêtes SPARQL et de visualiser les résultats via une interface de visualisation. Nous avons développé des interfaces de visualisation reposant sur LDViz pour permettre aux experts en génomique de blé d'exploiter et d'interagir avec le graphe WheatGenomicsSLKG.

3. Questions de Compétences

La littérature scientifique liées à la génomique fonctionnelle du blé est en constante évolution comme le montre la Figure 3.1. Son exploitation devrait aider les scientifiques à découvrir des interactions potentielles entre des entités d'intérêt en examinant leur co-occurrence dans les résumés des articles scientifiques. Dans le cadre de ce travail de recherche, les CQ ont permis d'éliciter les attentes des chercheurs travaillant sur la sélection génomique du blé et s'intéressant à explorer la littérature scientifique autour de ce sujet et à dévoiler de potentielles interactions entre des entités d'intérêt.

CQ1. Quels sont les gènes mentionnés à proximité d'un trait spécifique (par exemple, la résistance à la fusariose (resistance to Fusarium head blight), la résistance à la rouille des feuilles (resistance to leaf rust) ?

Cette question souligne l'importance d'assister les experts dans l'identification des entités génétiques potentiellement liées à un trait spécifique de la plante de blé, en vue

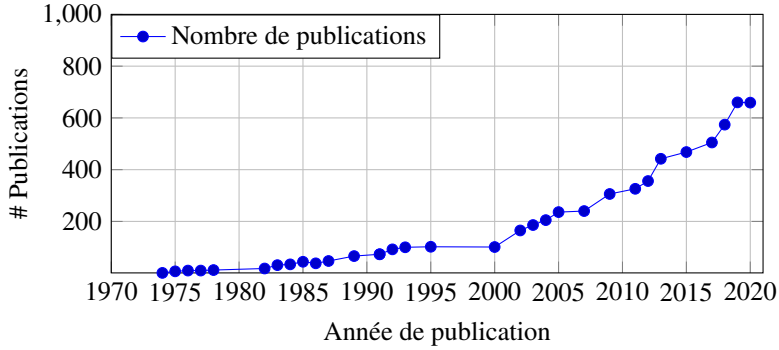
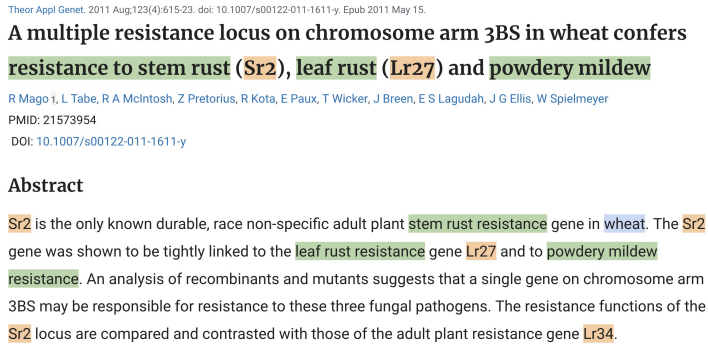


FIGURE 3.1 – Évolution du nombre d'articles publiés sur PubMed en relation avec la sélection assistée dans la culture de blé.

d'établir des liens éventuels entre les expressions génétiques et les traits observables. Ainsi, CQ1 permet la récupération de tous les contextes (publications) dans lesquels un trait spécifique apparaît, offrant ainsi un éclairage sur les autres entités qui sont présentes dans ce même contexte. Comme le montre la figure 3.2, ceci permet d'identifier les noms de gènes mentionnés à proximité d'un trait spécifique. La fréquence d'apparition d'un même gène à proximité d'un trait donne également des indications sur la corrélation entre le gène et le trait. Par exemple, en considérant la résistance à la rouille des feuilles, plusieurs gènes sont identifiés comme étant associés à ce trait dans plusieurs variétés de blé. Le gène *Lr34* apparaissant le plus fréquemment à proximité de ce trait est largement reconnu comme étant le principal gène impliqué dans la résistance à la rouille des feuilles. Ce type d'information est extrêmement précieux pour les programmes de sélection du blé, visant à développer des variétés résistantes à des maladies spécifiques.


Theor Appl Genet. 2011 Aug;123(4):615-23. doi: 10.1007/s00122-011-1611-y. Epub 2011 May 15.

A multiple resistance locus on chromosome arm 3BS in wheat confers resistance to stem rust (Sr2), leaf rust (Lr27) and powdery mildew

R Mago 1, L Tabe, R A McIntosh, Z Pretorius, R Kota, E Paux, T Wicker, J Breen, E S Lagudah, J G Ellis, W Spielmeyer

PMID: 21573954

DOI: 10.1007/s00122-011-1611-y

Abstract

Sr2 is the only known durable, race non-specific adult plant stem rust resistance gene in wheat. The Sr2 gene was shown to be tightly linked to the leaf rust resistance gene Lr27 and to powdery mildew resistance. An analysis of recombinants and mutants suggests that a single gene on chromosome arm 3BS may be responsible for resistance to these three fungal pathogens. The resistance functions of the Sr2 locus are compared and contrasted with those of the adult plant resistance gene Lr34.

FIGURE 3.2 – Exemple de reconnaissance d'entités nommées dans une publication PubMed

CQ2. Quels sont les marqueurs génétiques qui apparaissent à proximité d'un gène spécifique qui lui même est mentionné à proximité d'un trait de blé en particulier ?

L'objectif de la CQ2 est d'identifier des marqueurs candidats permettant de repérer les variétés de blé présentant un trait spécifique recherché. Les marqueurs génétiques, distinguant les différents allèles d'un gène grâce au polymorphisme de la séquence d'ADN, sont utilisés pour sélectionner les variétés de blé présentant un trait ou un phénotype d'intérêt agronomique. Par exemple, dans le cas de la *résistance à la maladie de la rouille jaune* dans la culture de blé, le gène *Yr65* est fréquemment cité dans la littérature à proximité de ce trait. D'autre part, des marqueurs tels que *Xgdm33*, *Xgwm11*, *Xgwm18* et *Xgwm413* sont mentionnés dans le même contexte que ce gène.

Étant donnée l'évolution des techniques de sélection des marqueurs génétiques au fil du temps et l'obsolescence de certaines d'entre elles, les experts peuvent souhaiter affiner leurs requêtes pour ne sélectionner que les publications postérieures à 2010. Cette question de compétence souligne ainsi l'importance d'intégrer dans le graphe des méta-données descriptives des publications, telles que l'année de publication, la liste des auteurs ou encore le nombre de citations.

CQ3. Quelles publications scientifiques mentionnent des gènes à proximité d'un nom de variété spécifique de blé (i.e. Arina*) et d'un trait d'une classe de traits (par exemple, tous les traits liés à la résistance aux pathogènes fongiques) ?*

La question CQ3 met en évidence la nécessité pour les experts d'effectuer une analyse documentaire systématique des publications mentionnant des gènes spécifiques mentionnés dans la littérature en relation avec certains traits, ainsi que des variétés de blé. Les résultats de cette recherche devraient comprendre une liste d'articles mentionnant, dans leur résumé ou leur titre, des noms de gènes, une variété de blé et un ensemble de traits connus pour être impliqués dans la résistance aux agents pathogènes, à titre d'exemple. Par exemple, un expert pourrait s'intéresser à la résistance aux champignons pathogènes responsables de pertes massives et dévastatrices pour les cultures. L'étude des mécanismes de résistance est donc essentielle pour une compréhension approfondie des interactions entre les agents pathogènes et les variétés de blé. En se basant sur la structure hiérarchique de l'ontologie WTO - Wheat Trait and Phenotype Ontology [28], cette CQ nécessite d'inclure tous les traits appartenant à la classe de résistance aux pathogènes fongiques. Cette question souligne l'importance d'intégrer les connaissances du domaine, formellement représentées dans les ressources ontologiques et terminologiques telles que WTO pour annoter les EN dans le texte.

CQ4 : Quels sont les noms de gènes qui sont cités dans la littérature à proximité d'un taxon spécifique (et éventuellement d'un ou de plusieurs de ses descendants) ?

La question CQ4 reflète la nécessité d'effectuer une recherche sur les mentions de gènes cités à proximité de différentes mentions de taxons. Nous pouvons lancer la requête en nous concentrant sur une seule mention de taxon et l'étendre dynamiquement en incluant chaque taxon descendant. Ainsi, la requête affiche les premiers résultats d'une recherche unique sur une mention de taxon spécifique. Elle génère ensuite un

ensemble plus complet de résultats. En conclusion, la Figure 3.3 résume les différentes interactions potentielles entre les entités du graphe WheatGenomics-SLKG. Ces interactions sont reflétées par différentes CQs mentionnées précédemment. Ces interactions ne sont pas explicitement décrites dans le graphe mais peuvent être identifiées par les experts lors de leur exploration.

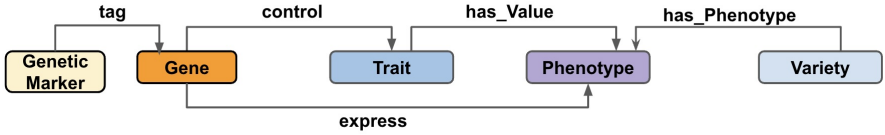


FIGURE 3.3 – Les classes d'entités nommées dans le graphe WheatGenomics-SLKG et leurs potentielles relations sémantiques

4. MODÈLE PROPOSÉ

La définition d'un modèle qui capture la nature des entités et leurs relations dans le graphe de connaissances est impérative. En effet, le futur graphe de connaissances intégrera différent types d'entités dont la sémantique sera décrite différemment selon la nature de l'entité. Nous avons d'une part les publications scientifiques et d'autres part les entités nommées. Nous réutilisons un ensemble de vocabulaires largement répandus, adaptés à la représentation des articles et des entités nommées en RDF. La partie centrale de ce modèle s'appuie sur le modèle précédemment proposé dans [25] et l'étend. Il est basé sur l'ontologie d'annotation Web (Ontology Web Annotation – OA) du W3C [34] pour structurer, décrire et intégrer les EN extraites du corpus, et sur cinq vocabulaires complémentaires pour décrire les articles scientifiques du corpus. Le tableau 4.1 montre les principaux vocabulaires utilisés pour décrire les annotations d'entités nommées ainsi que les documents scientifiques.

Prefix	**Namespace**
oa	`http://www.w3.org/ns/oa#`
dct	`http://purl.org/dc/terms/`
dce	`http://purl.org/dc/elements/1.1/`
fabio	`http://purl.org/spar/fabio/`
bibo	`http://purl.org/ontology/bibo/`
schema	`http://schema.org/`
frbr	`http://purl.org/vocab/frbr/core#`

TABLE 4.1 – Liste des vocabulaires utilisés dans le graphe WheatGenomics-SLKG

4.1. Description des Entités Nommées en se basant sur OA

OA est une ontologie permettant de structurer et de partager tout type d'annotations dans un format interopérable [34]. Le classe de base de l'ontologie OA est la classe `oa:Annotation`, Une instance a_i de cette classe représente l'occurrence d'une mention m_e d'une entité nommée e dans le titre ou le résumé (ou l'une de ses sous-sections) d'un article à une position de début d et une position de fin f. Pour formaliser cette annotation, nous disposons dans l'ontologie de plusieurs propriétés permettant de décrire le contexte d'apparition de cette EN.

- La propriété `oa:hasTarget` identifie la partie du document qui est annotée avec l'annotation a_i. La cible est une sélection de ressources avec un sélecteur, c'est-à-dire une ressource qui identifie la partie du texte m_e qui mentionne une entité reconnue e. Dans ce travail, nous utilisons deux types de sélecteurs : `oa:TextQuoteSelector` et `oa:TextPositionSelector` pour indiquer respectivement la mention m_e de l'EN, la position de début et de fin de m_e dans le texte. La propriété `oa:hasSource` est utilisée pour spécifier l'URI de la source où le sélecteur est appliqué, la source étant soit l'URI du document, soit l'une de ses sous-parties (résumé/titre).
- a_i a un corps `oa:hasBody` qui renvoie vers l'URI d'une entité e définie dans une ontologie ou un vocabulaire existant du domaine tels par exemple l'URI d'un concept WTO [28] ou d'une classe de taxons dans la taxonomie NCBI[(10)].

La figure 4.1 illustre un exemple de graphe RDF qui capture cinq instances d'annotations NE reconnues dans le titre et le résumé d'une publication dans le corpus PubMed[(11)]. Le titre et le résumé de la publication sont identifiés par un URI et deviennent la source du sélecteur de cible. Trois annotations sont liées avec la propriété `oa:hasBody` à des concepts SKOS de la ressource WTO (zone jaune dans la figure 4.1). Toutes les mentions d'EN sont identifiées par deux sélecteurs :

- une instance de `oa:TextQuoteSelector` est utilisée pour spécifier le texte de la mention.
- une instance de `oa:TextPositionSelector` est utilisée pour spécifier les positions de début et de fin de la mention.

[(10)]NCBI Taxonomy `https://www.ncbi.nlm.nih.gov/taxonomy`
[(11)]`https://pubmed.ncbi.nlm.nih.gov/21573954/`

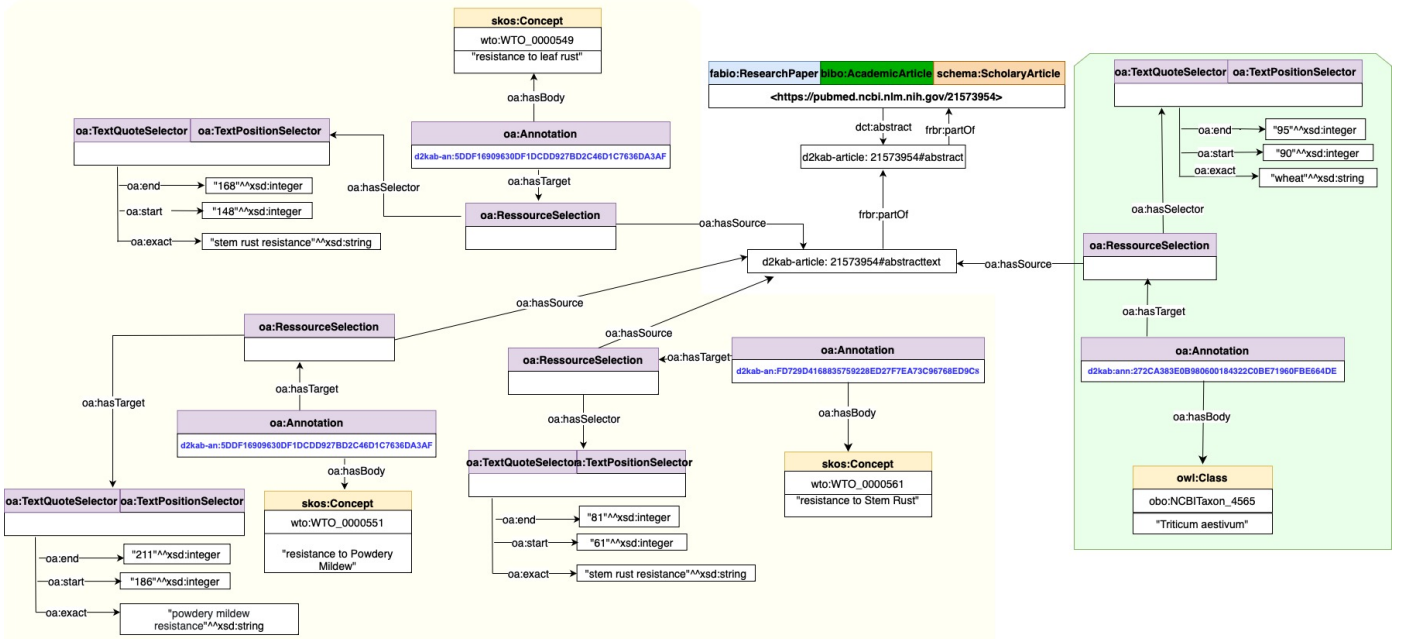


FIGURE 4.1 – Exemple de graphe RDF représentant une publication du corpus

4.2. Description des articles scientifiques

Pour représenter et décrire les publications scientifiques, nous avons réutilisé les vocabulaires suivants : Dublin Core, FRBR aligned Bibliographic Ontology (FaBiO), Bibliographic Ontology (BIBO) et Schema.org. Ces vocabulaires définissent une liste exhaustive de méta-données descriptives pour décrire les publications scientifiques telles que le DOI, l'année de publication, le nombre de pages, le journal, etc. Ainsi, un article scientifique sera représenté comme une instance des classes `fabio:ResearchPaper` et `bibo:AcademicArticle`. Les propriétés Dublin Core `dct:title` et `dct:abstract` permettent de relier la publication à son titre et son résumé. Certains résumés d'articles scientifiques sont structurés en 3 sous-sections que nous représentons comme étant 3 entités différentes identifiées chacune par un URI unique. La propriété `frbr:partOf` sera utilisée pour représenter le lien *partie de* entre un résumé et une publication et aussi entre un résumé et ses sous-sections. La figure 4.2 illustre un graphe RDF représentant une publication scientifique avec un sous-ensemble de ses méta-données descriptives, à savoir le titre de la publication, le résumé et ses sous-sections (*background*, *results*, *conclusions*). La figure 4.2 illustre (un sous-ensemble) des métadonnées bibliographiques d'un document scientifique.

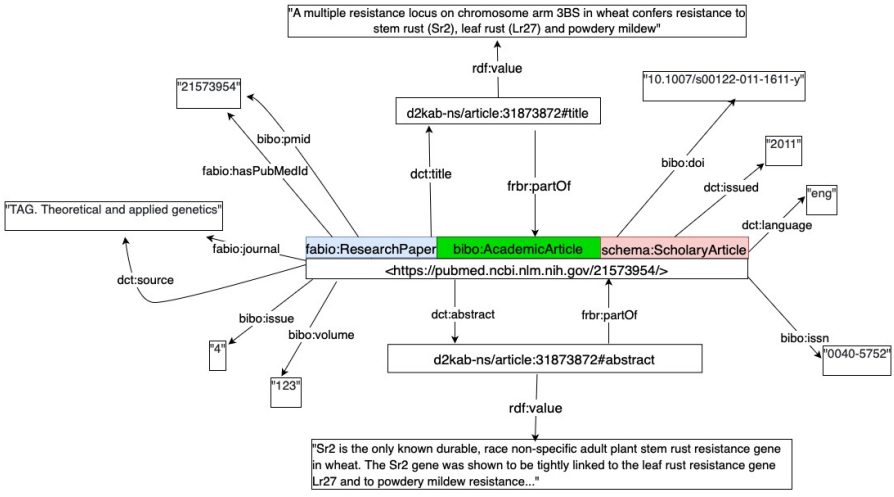


Figure 4.2 – Exemple de graphe RDF représentant une publication du corpus

5. Construction du graphe de connaissances

Nous présentons dans cette section le jeu de données (corpus) et les outils utilisés pour la construction du graphe de connaissances.

5.1. Jeu de données

Le groupe de recherche MaIAGE a rassemblé 8 496 références scientifiques d'articles publiés entre 1974 et 2021 portant sur les techniques modernes de phénotypage et de dépistage génomique pour la sélection ciblée du blé. Une première exploration du corpus montre qu'au cours des deux dernières décennies, le nombre de publications a significativement augmenté. Comme le montre la figure 3.1, tandis que la première publication est apparue en 1974, le nombre de publications par année a augmenté lentement entre 1975 et 2000. Plus de 80 % des publications du corpus proviennent des deux dernières décennies, ce qui reflète l'intérêt porté par la communauté de recherche au développement de nouvelles techniques de sélection génomique de nouvelles variétés de blé avec de traits agronomiques intéressants.

Étant donné le corpus des 8 496 publications, un ensemble de 4318 entités nommées a été extrait en utilisant la plate-forme AlvisNLP. AlvisNLP offre une chaîne de TAL pour l'annotation sémantique de documents textuels dans le domaine de la génomique des plantes. Elle intègre plusieurs outils permettant la segmentation du texte en mots/phrases, la reconnaissance d'entités nommées, l'analyse de termes, le typage sémantique et l'extraction de relations présentes entre entités.

Le format du jeu de données fourni par l'équipe MaIAGE comprend différent types d'informations stockées séparément dans plusieurs fichiers CSV. Pour chaque publication du corpus, l'identifiant PubMed, le titre, le résumé et les possibles sous-sections du résumé sont fournis. Les entités nommées extraites par AlvisNLP ont été stockées dans un autre fichier CSV. Pour chaque occurrence d'entité, plusieurs informations sont renseignées sur plusieurs colonnes : l'identifiant Pubmed de l'article, la section du résumé où apparaît cette occurrence, la classe assignée à cette entité (gène, trait, phénotype, marqueur, variété, taxon), la position (offset) de début et de fin de la mention de l'entité dans le texte. Enfin, les relations détectées par AlvisNLP sont stockées dans un troisième fichier CSV. Dans ce jeu, nous avons uniquement la relation *variety_has_phenotype* qui permet de relier une occurrence d'une entité nommée de type variété à une occurrence d'une entité nommée de type phénotype.

	Template d'URI
Entity	http://ns.inria.fr/d2kab/{EntityClass}/{EntityID}
Article	http://ns.inria.fr/d2kab/article/{PubmedId}
Annotation	http://ns.inria.fr/d2kab/annotation/{annotationId}
Title	http://ns.inria.fr/d2kab/article/{PubmedId}#title
Abstract	http://ns.inria.fr/d2kab/article/{PubmedId}#abstract
Abstract section	http://ns.inria.fr/d2kab/article/{PubmedId} #{sectionName}
Relation	http://ns.inria.fr/d2kab/relation/{relationId}

Table 5.1 – Template de génération d'URI des ressources de notre graphe

5.2. Transformation en RDF

Pour transformer les annotations produites par AlvisNLP en un graphe RDF, nous avons utilisé l'outil Morph-xR2RML [21, 27]. Tout d'abord, nous avons défini un ensemble de règles de mapping. Ces règles décrivent un ensemble de *TripleMap* respectant le vocabulaire et la syntaxe qui sont fournis par la spécification xR2RML[(12)]. Chaque *TripleMap* définit un patron générique pour la génération de triplets RDF en respectant la modélisation proposée dans la section 4. L'ensemble des règles xR2RML définies sont disponibles dans le répertoire GitHub du projet[(13)]. La table 5.1 décrit les patrons pour génération d'URI des différentes ressources du futur graphe. De plus, dans l'objectif d'enrichir le graphe avec des méta-données descriptives des publications scientifiques, nous nous sommes appuyés sur l'architecture de micro-services SPARQL [24], et avons défini un micro-service[(14)] permettant d'interroger l'API PubMed Central et obtenir les méta-données (représentées en RDF) de chaque publication étant donné l'identifiant *PubMed* de la publication. Cette représentation RDF se base sur la modélisation présentée dans la section 4.2. Le tableau 5.2 représente le nombre de triplets pour chaque classe du graphe de connaissances.

Classe	
Nbre total d'annotations	88 880
Nbre total d'articles	8 496
Nbre total de gènes	1 160
Nbre total de taxons	2 462
Nbre total de traits	98
Nbre total de marqueurs	521
Nbre total de variétés	77
Nbre total de relations	162

Table 5.2 – Nombre d'entités par classes dans notre graphe de connaissances

6. Publication et Interrogation du Graphe

Le graphe RDF WheatGenomics-SLKG est identifié via un DOI, téléchargeable depuis la plateforme Zenodo[(15)] et accessible au moyen d'un endpoint SPARQL public[(16)]. Le dépôt Github[(17)] du projet fournit une documentation exhaustive incluant des détails relatifs aux représentations RDF, aux graphes nommés et aux ontologies chargées dans l'endpoint. Conformément aux principes de la science ouverte, tous les scripts, fichiers de configuration et de traduction en RDF impliqués dans notre pipeline sont fournis dans le dépôt Github du projet selon les termes de la licence Apache 2.0.

(12) https://www.i3s.unice.fr/~fmichel/xr2rml_specification_v5.html
(13) https://github.com/Wimmics/WheatGenomicsSLKG/tree/main/mapping-rules
(14) https://sparql-micro-services.org/service/pubmed/getArticleByPMId_sd/
(15) https://zenodo.org/records/10410742
(16) http://d2kab.i3s.unice.fr/sparql
(17) https://github.com/Wimmics/WheatGenomicsSLKG/

Les CQ définies par l'expert du domaine et présentées dans la section 3 ont été traduites en requêtes SPARQL[18] et leurs résultats ont été validés. Nous nous contentons de présenter dans cette section les requêtes SPARQL implémentant les questions de compétence *CQ1* et *CQ3* présentées respectivement dans les listings 1 et 2. La requête SPARQL présentée dans le listing 1 implémente la CQ1 et permet aux scientifiques d'identifier les gènes qui sont mentionnés à proximité du trait *resistance to leaf rust*. De plus, la requête calcule le nombre de fois qu'un gène et le trait en question sont identifiés dans un même contexte. Les résultats de cette requête confirment que Lr34 est le gène le plus cité dans la littérature. Les gènes Lr10, Lr26 et Lr24 apparaissent également comme les gènes les plus fréquents.

Listing 1 – Requête SPARQL implémentant la question de compétence CQ1

```
SELECT ?GeneName (count(distinct ?paper) as ?NbOcc)
FROM NAMED <http://ns.inria.fr/d2kab/graph/wheatgenomicsslkg>
FROM NAMED <http://ns.inria.fr/d2kab/ontology/wto/v3>
WHERE {
GRAPH <http://ns.inria.fr/d2kab/graph/wheatgenomicsslkg> {
  ?a1 a oa:Annotation;
      oa:hasTarget [ oa:hasSource ?source1 ]~;
      oa:hasBody ?WTOtraitURI .

  ?source1 frbr:partOf+ ?paper .

   ?a a oa:Annotation~;
      oa:hasTarget [ oa:hasSource ?source ]~;
      oa:hasBody [ a d2kab:Gene; skos:prefLabel ?GeneName ].

   ?source frbr:partOf+ ?paper.

   ?paper a fabio:ResearchPaper.
}
GRAPH <http://ns.inria.fr/d2kab/ontology/wto/v3> {
   ?WTOtraitURI skos:prefLabel "resistance to Leaf rust" .
}
}
GROUP BY ?GeneName
HAVING (count(distinct ?paper) > 1)
ORDER BY DESC(?NbOcc)
```

La requête SPARQL, présentée dans le listing 2, implémente la CQ3 et permet aux scientifiques d'extraire les publications dans lesquelles les gènes sont mentionnés à proximité des variétés de blé et des traits d'une classe spécifique, par exemple, tous les traits du blé liés à la résistance aux pathogènes fongiques. Basée sur la structure de WTO qui classe les traits dans différentes taxonomies, la requête extrait tous les caractères appartenant à la sous-hiérarchie de la classe de résistance aux pathogènes fongiques.

[18] https://github.com/Wimmics/WheatGenomicsSLKG/tree/main/sparql-queries

Listing 2 – Requête SPARQL implémentant la question de compétence CQ3

```
SELECT distinct ?paper ?Title ?GeneName ?varietyName ?WTOtrait
FROM NAMED <http://ns.inria.fr/d2kab/graph/wheatgenomicsslkg>
FROM NAMED <http://ns.inria.fr/d2kab/ontology/wto/v3>
WHERE {
  GRAPH <http://ns.inria.fr/d2kab/graph/wheatgenomicsslkg> {
    ?a1 a oa:Annotation~;
       oa:hasTarget [ oa:hasSource ?source1 ]~;
       oa:hasBody [ a d2kab:Gene; skos:prefLabel ?GeneName ] .

    ?source1 frbr:partOf+ ?paper .

    ?a2 a oa:Annotation~;
       oa:hasTarget [ oa:hasSource ?source2 ]~;
       oa:hasBody ?body .
    ?source2 frbr:partOf+ ?paper .

    ?a3 a oa:Annotation~;
       oa:hasTarget [ oa:hasSource ?source3 ]~;
       oa:hasBody [ a d2kab:Variety; skos:prefLabel ?varietyName ] .
    ?source3 frbr:partOf+ ?paper .

    ?paper a fabio:ResearchPaper~; dct:title ?titleURI .
    ?titleURI rdf:value ?Title .
    }
  GRAPH <http://ns.inria.fr/d2kab/ontology/wto/v3> {
    {
    ?body skos:prefLabel ?WTOtrait~; a ?class .
    ?class rdfs:subClassOf* <http://opendata.inrae.fr/wto/0000246> .
    }
    UNION
    {
    ?body rdfs:label ?WTOtrait~;
         rdfs:subClassOf* <http://opendata.inrae.fr/wto/0000246> .
    }
    UNION
    {
    ?body skos:prefLabel ?WTOtrait~; skos:broader* ?concept .
    ?concept a ?class .
    ?class rdfs:subClassOf* <http://opendata.inrae.fr/wto/0000246> . }
    }
}
LIMIT 20
```

7. Visualisation et utilisations du graphe de connaissances

Afin de permettre à tout utilisateur, non expert des graphes de connaissances, d'exploiter et interroger le graphe SLKG pour la génomique de blé, nous avons développé plusieurs interfaces web.

7.1. Recherche par co-occurrence d'entités nommées

Une première interface permet la recherche d'articles mentionnant une ou plusieurs entités nommées, quels que soient leur type (gène, marqueur, taxon, trait ou phénotype)[(19)]. La saisie de chaque entité se fait en entrant quelques lettres, l'auto-complétion

[(19)]Cette interface est actuellement disponible à l'URL `http://d2kab.i3s.unice.fr/wheatgenomics/search/`.

propose les labels des entités correspondants, labels principaux ou alternatifs, et l'utilisateur sélectionne les entités qui correspondent à sa recherche.

Cette interface peut aider à répondre à certaines questions de compétence, comme la CQ3 illustrée dans la Figure 7.1 : l'utilisateur a saisi la variété *Arina* et le trait de *résistance aux pathogènes fongiques* (haut de la figure). Les résultats consistent en deux sections. La première section donne les articles mentionnant exactement les entités nommées sélectionnées. Dans l'exemple, celle-ci est vide. La seconde section liste les articles mentionnant soit les entités nommées sélectionnées soit des entités plus spécifiques (sous-concepts ou sous-classes). Dans l'exemple, deux articles sont annotés avec la variété *Arina* et le trait *résistance à la rouille*, une sous-classe de la *résistance aux pathogènes fongiques*. En cliquant sur l'icône à droite de chaque résultat, l'utilisateur ouvre une nouvelle page permettant de visualiser les entités nommées du résumé, comme illustré dans la Figure 7.2. Parmi ces entités nommées on trouve plusieurs gènes, ce qui étaient l'objet de la CQ3.

Cette interface se compose de deux projets (application web front-end et services backend) publiés sous licence ouverte et chacun identifié par un DOI [22, 23].

7.2. Exploitation du graphe des entités nommées

Une deuxième interface est destinée à offrir à l'utilisateur une approche visuelle plus adaptée à la résolution de questions de compétences complexes. Elle s'appuie sur l'outil LDViz [20] et permet l'exploration du WheatGenomics-SLKG à travers la co-occurrence d'entités, celles-ci pouvant être des entités nommées mentionnées dans les articles, des auteurs, des institutions etc.

La Figure 7.3 illustre l'utilisation de cette interface pour répondre à la question de compétence CQ1 : découvrir les gènes mentionnés à proximité du trait de la *résistance à la rouille*. Le processus se déroule comme suit :

- L'exploration débute par la sélection d'un type de requête pré-configuré et la saisie du concept, ici *resistance to rust* (1).
- LDViz exécute la requête puis présente les résultats en (2). Les noeuds verts représentent la *résistance à la rouille* ou un de ses sous-concepts ; les noeuds oranges représentent des gènes. Un lien entre deux noeuds indique qu'ils co-occurrent dans au moins une publication.
- L'utilisateur clique-droit sur la résistance à la rouille, et sélectionne la vue égocentrique (3). Dans celle-ci, la résistance à la rouille est entourée de chacun des gènes co-occurrents, la taille de la barre bleue indique le nombre de publications ayant cette co-occurrence.
- L'utilisateur clique-droit sur le gène Lr10 et ouvre la liste des publications relatives (4) et leur distribution dans le temps (5).

Notons que la flexibilité de LDViz permet d'utiliser ces mêmes visualisations pour des entités de toutes natures. Dans la Figure 7.4 par exemple, l'utilisateur explore le réseau des co-auteurs du corpus. Le graphe visualisé en (2) représente cette fois-ci des

FIGURE 7.1 – Recherche des articles mentionnant à la fois la variété Arina et un trait de type résistance aux pathogènes fongiques.

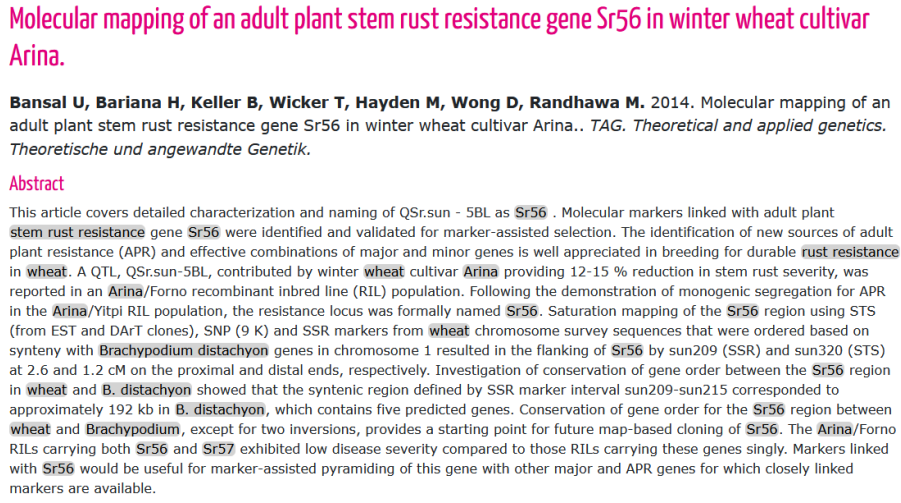


FIGURE 7.2 – Visualisation des entités nommées du résumé d'un article.

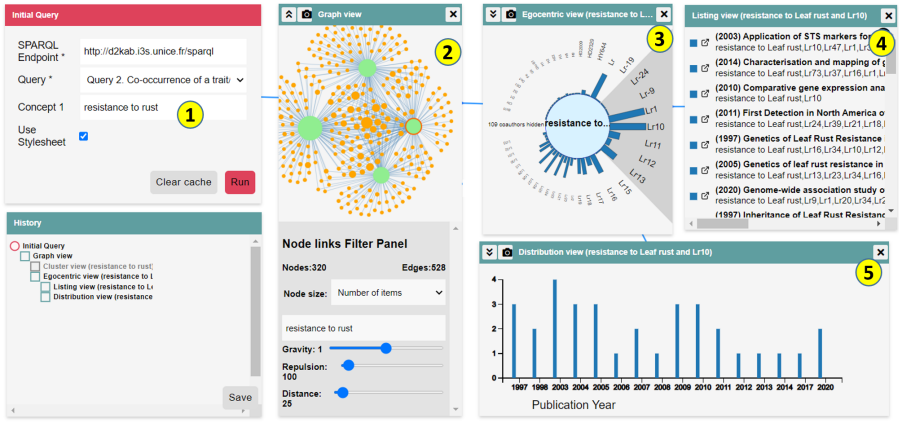


FIGURE 7.3 – Utilisation de LDViz pour découvrir les gènes mentionnés à proximité du trait de la *résistance à la rouille* et les publications associées.

auteurs. Un lien entre deux auteurs signifie qu'ils ont co-écrit au moins une publication. La vue suivante (3) est cette fois-ci une vue de type clusters : elle illustre les groupes de co-auteurs parmi tous les auteurs ayant co-écrit avec Dubcovsky. Ici on s'intéresse au cluster Dubcovsky, Chao, Anderson, et les publications qu'ils ont co-écrites sont listées en (4).

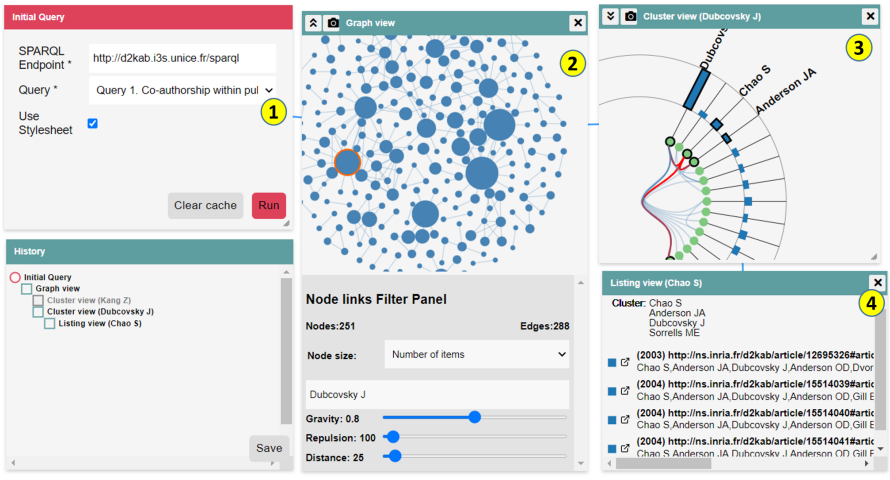


FIGURE 7.4 – Utilisation de LDViz pour explorer le réseau des co-auteurs du corpus.

8. Conclusion et Travaux futurs

Explorer la littérature scientifique en rapport avec les concepts clés la génomique des plantes peut s'avérer une tâche ardue pour les chercheurs. Ce travail de recherche s'attaque aux problèmes d'une recherche bibliographique transversale, et du liage des informations extraites à partir de la littérature scientifique avec les bases de données génomiques. En effet, la disponibilité de ressources sémantiques dans ce domaine (ontologies, thésaurus) peut s'avérer d'une grande utilité pour annoter les textes scientifiques et extraire les entités nommées. Dans ce papier, nous avons conçu et construit un graphe de connaissances WheatGenomics-SLKG en considérant les annotations extraites à partir d'un corpus de publications scientifiques. Ces annotations sont produites par la plate-forme AlvisNLP et portent sur différentes entités nommées de différents types incluant des gènes, des phénotypes, des marqueurs génétiques, des variétés et des taxons en rapport avec la génomique du blé. Dans WheatGenomics-SLKG, les contextes d'apparition des différentes entités sont décrits et représentés d'une manière structurée en se basant sur l'utilisation conjointe des vocabulaires standards du Web sémantique (*i.e.*, [34]) et des ontologies du domaine en question (*i.e.*, [28]). Ceci a permis de les interroger de manière uniforme avec le langage SPARQL et surtout d'exploiter les contextes d'apparition pour découvrir des associations implicites entre ces entités. Comme travaux futurs, nous envisageons d'intégrer dans le graphe de connaissances des observations collectées par des professionnels du domaine. Ces observations décrivent les stades de croissance des plantes, la fréquence d'attaque de maladies pour certaines variétés, les localisations géographiques des parcelles de culture, les paramètres météorologiques, etc. L'objectif serait de permettre le développement de modèles combinant des connaissances émanant de la littérature scientifique et des données d'observations.

9. Remerciements

Ce travail a été réalisé dans le cadre du projet « Des Données aux Connaissances en Agronomie et Biodiversité » (D2KAB–www.d2kab.org) financé par l'Agence Nationale de la Recherche (ANR-18-CE23-0017).

Bibliographie

[1] M. Ba & R. Bossy, « Interoperability of corpus processing work-flow engines: the case of AlvisNLP/ML in OpenMinTeD », in *Proceedings of the Workshop on Cross-Platform Text Mining and Natural Language Processing Interoperability (INTEROP 2016) at LREC 2016* (Portorož, Slovenia), 2016, p. 15-18.

[2] E. Bernasconi, M. Ceriani, D. Di Pierro, S. Ferilli & D. Redavid, « Linked Data Interfaces: A Survey », *Inf.* **14** (2023), nº 9, p. 483.

[3] R. Bossy, L. Deleger, E. Chaix, M. Ba & C. Nédellec, « Bacteria Biotope at BioNLP Open Shared Tasks 2019 », in *5th Workshop on BioNLP Open Shared Tasks BioNLP-OSTEMNLP-IJCNLP 2019, ACL*, 2019.

[4] J. Chen, H. Dong, J. Hastings, E. Jiménez-Ruiz, V. López, P. Monnin, C. Pesquita, P. Škoda & V. Tamma, « Knowledge Graphs for the Life Sciences : Recent Developments, Challenges and Opportunities », *Transactions on Graph Data and Knowledge* **1** (2023), nº 1, article no. 5 (33 pages).

[5] I. Davis & R. Newman, « Expression of Core FRBR Concepts in RDF », `https://vocab.org/frbr/core`.

[6] P.-Y. Genest, P.-E. Portier, E. Egyed-Zsigmond & L.-W. Goix, « PromptORE – A Novel Approach Towards Fully Unsupervised Relation Extraction », in *Proceedings of the 31st ACM International Conference on Information & Knowledge Management*, ACM, 2022, p. 561-571.

[7] J. M. Giorgi & G. D. Bader, « Towards reliable named entity recognition in the biomedical domain », *Bioinformatics* **36** (2019), nº 1, p. 280-286, `https://arxiv.org/abs/https://academic.oup.com/bioinformatics/article-pdf/36/1/280/31813710/btz504.pdf`.

[8] M. Hofer, D. Obraczka, A. Saeedi, H. Köpcke & E. Rahm, « Construction of Knowledge Graphs: Current State and Challenges », *Information* **15** (2024), nº 8, article no. 509 (61 pages).

[9] A. Hogan, E. Blomqvist, M. Cochez, C. d'Amato, G. de Melo, C. Gutierrez, J. E. Labra Gayo, S. Kirrane, S. Neumaier, A. Polleres, R. Navigli, A. Ngonga Ngomo, S. M. Rashid, A. Rula, L. Schmelzeisen, S. J. F., S. Staab & A. Zimmermann, « Knowledge Graphs », `https://arxiv.org/abs/2003.02320`, 2020.

[10] M.-S. Huang, J.-C. Han, P.-Y. Lin, Y.-T. You, R. T.-H. Tsai & W.-L. Hsu, « Surveying biomedical relation extraction: a critical examination of current datasets and the proposal of a new resource », *Briefings in Bioinformatics* **25** (2024), nº 3, article no. bbae132 (17 pages).

[11] P.-L. Huguet Cabot & R. Navigli, « REBEL: Relation Extraction By End-to-end Language generation », in *Findings of the Association for Computational Linguistics: EMNLP 2021* (Punta Cana, Dominican Republic) (M.-F. Moens, X. Huang, L. Specia & S. W.-t. Yih, éds.), Association for Computational Linguistics, 2021, p. 2370-2381.

[12] E. Ikkala, E. Hyvönen, H. Rantala & M. Koho, « Sampo-UI: A full stack JavaScript framework for developing semantic portal user interfaces », *Semantic Web* **13** (2021), p. 69-84.

[13] J.-D. Kim, J.-J. Kim, X. Han & D. Rebholz-Schuhman, « Extending the evaluation of Genia Event task toward knowledge base construction and comparison to Gene Regulation Ontology task », *BMC bioinformatics* **16** (2015), p. S3.

[14] J.-D. Kima, K. Verspoorb, M. Dumontierc & K. B. Cohend, « Semantic representation of annotation involving texts and linked data resources », Semantic Web journal, 2015.

[15] P. Larmande & K. Todorov, « AgroLD: A knowledge graph for the plant sciences », in *ISWC 2021 – 20th International Semantic Web Conference* (Virtual, France), Lecture Notes in Computer Science, vol. 12922, Springer International Publishing, 2021, p. 496-510.

[16] R. Leaman, R. Khare & Z. Lu, « Challenges in clinical natural language processing for automated disorder normalization », *Journal of biomedical informatics* **57** (2015), p. 28-37.

[17] X. Lin, T. Liu, W. Jia & Z. Gong, « Distantly Supervised Relation Extraction using Multi-Layer Revision Network and Confidence-based Multi-Instance Learning », in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (Punta Cana, Dominican Republic) (M.-F. Moens, X. Huang, L. Specia & S. W.-t. Yih, éds.), Association for Computational Linguistics, 2021, p. 165-174.

[18] W. Liu, P. Zhou, Z. Zhao, Z. Wang, Q. Ju, H. Deng & P. Wang, « K-BERT: Enabling Language Representation with Knowledge Graph », `https://arxiv.org/abs/1909.07606`, 2019.

[19] S. Lohmann, V. Link, E. Marbach & S. Negru, « WebVOWL: Web-based Visualization of Ontologies », in *International Conference Knowledge Engineering and Knowledge Management* (Berlin, Heidelberg), Springer-Verlag, 2014, p. 154-158.

[20] A. Menin, P. Maillot, C. Faron, O. Corby, C. M. Dal Sasso Freitas, F. Gandon & M. Winckler, « LDViz: a tool to assist the multidimensional exploration of SPARQL endpoints », in *Web Information Systems and Technologies: 16th International Conference, WEBIST 2020*, LNBIP - Lecture Notes in Business Information Processing, vol. LNBIP – 469, Springer, 2023, p. 149-173.

[21] F. Michel, L. Djimenou, Faron-ZuckerCatherine & J. Montagnat, « Translation of Relational and Non-relational Databases into RDF with xR2RML », in *WEBIST 2015 – Proceedings of the 11th International Conference on Web Information Systems and Technologies, Lisbon, Portugal, 20-22 May, 2015* (V. Monfort, K. Krempels, T. A. Majchrzak & Z. Turk, éds.), SciTePress, 2015, p. 443-454.

[22] F. Michel & S. Essam, « WheatGenomicsSLKG Web Application Backend Services », `https://github.com/Wimmics/wheatgenomicsslkg-web-backend/tree/1.0`, 2024, [Software] V1.0, DOI : 10.5281/zenodo.10514504.

[23] ———, « WheatGenomicsSLKG Web Visualization », `https://github.com/Wimmics/wheatgenomicsslkg-web-visualization/tree/1.0`, 2024, [Software] V1.0, DOI: 10.5281/zenodo.10514502.

[24] F. Michel, C. Faron Zucker, O. Gargominy & F. Gandon, « Integration of Web APIs and Linked Data Using SPARQL Micro-Services – Application to Biodiversity Use Cases », *Information* **9** (2018), nº 12, article no. 310.

[25] F. Michel, F. Gandon, V. Ah-Kane, A. Bobasheva, E. Cabrio, O. Corby, R. Gazzotti, A. Giboin, S. Marro, T. Mayer, M. Simon, S. Villata & M. Winckler, « Covid-on-the-Web: Knowledge Graph and Services to Advance COVID-19 Research », in *19th International Semantic Web Conference, Athens, Greece, November 2-6, 2020, Proceedings, Part II*, Lecture Notes in Computer Science, vol. 12507, Springer, 2020, p. 294-310.

[26] F. Michel, F. Gandon, V. Ah-Kane, A. Bobasheva et al., « Covid-on-the-Web : Graphe de Connaissances et Services pour faire Progresser la Recherche sur la COVID-19 », in *IC 2021 – 32e Journées francophones d'Ingénierie des Connaissances* (Bordeaux, France), Maxime Lefrançois, 2021, p. 1-9.

[27] F. Michel, F. Priyatna & O. Corcho, « Morph-xR2RML », `https://github.com/frmichel/morph-xr2rml/tree/morph-xr2rml-1.3.2`, 2021, [Software] V1.3.2, SWHID: `swh:1:rev:2494b1da7b128e38edc7759f090201030c64211b`.

[28] C. Nédellec, L. Ibanescu, R. Bossy & P. Sourdille, « WTO, an ontology for wheat traits and phenotypes in scientific publications », *Genomics & Informatics* **18** (2020), article no. e14.

[29] C. Nédellec, C. Sauvion, R. Bossy, M. Borovikova & L. Deléger, « TaeC: A manually annotated text dataset for trait and phenotype extraction and entity linking in wheat breeding literature – journals.plos.org », *PLoS ONE* **19** (2024), nº 6, article no. e0305475.

[30] N. Perera, M. Dehmer & F. Emmert-Streib, « Named Entity Recognition and Relation Detection for Biomedical Information Extraction », *Frontiers in Cell and Developmental Biology* **8** (2020), article no. 673.

[31] S. Peroni & D. Shotton, « FaBiO and CiTO : Ontologies for describing bibliographic resources and citations », *Journal of Web Semantics* **17** (2012), p. 33-43.

[32] A. Randles, L. McKenna, L. Kilgallon, B. Yaman, P. Crooks & D. O'Sullivan, « The Knowledge Graph Explorer for the Virtual Record Treasury of Ireland », in *Proceedings of the 9th International Workshop on the Visualization and Interaction for Ontologies, Linked Data and Knowledge Graphs co-located with the 23rd International Semantic Web Conference (ISWC 2024), Baltimore, USA, November 12, 2024* (B. Fu, P. Lambrix, H. Li, S. Nunes & C. Pesquita, éds.), CEUR Workshop Proceedings, vol. 3773, CEUR-WS.org, 2024, p. 47-61.

[33] K. E. Ravikumar, M. Rastegar-Mojarad & H. Liu, « BELMiner: adapting a rule-based relation extraction system to extract biological expression language statements from bio-medical literature evidence sentences », *Database* **2017** (2017), article no. baw156 (12 pages).

[34] R. Sanderson, P. Ciccarese & B. Young, « Web Annotation Ontology », `https://www.w3.org/TR/annotation-vocab/`, 2017.

[35] R. Usbeck, A.-C. Ngonga Ngomo, M. Röder, D. Gerber, S. Coelho, S. Auer & A. Both, « AGDISTIS – Graph-Based Disambiguation of Named Entities Using Linked Data », in *International Workshop on the Semantic Web* (Cham), Springer International Publishing, 2014, p. 457-471.

[36] M. Weise, S. Lohmann & F. Haag, « LD-VOWL: Extracting and Visualizing Schema Information for Linked Data Endpoints », in *Proceedings of the Second International Workshop on Visualization and Interaction for Ontologies and Linked Data co-located with the 15th International Semantic Web Conference, VOILAISWC 2016, Kobe, Japan, October 17, 2016* (V. Ivanova, P. Lambrix, S. Lohmann & C. Pesquita, éds.), CEUR Workshop Proceedings, vol. 1704, CEUR-WS.org, 2016, p. 120-127.

[37] N. Yacoubi, D. Graux & C. Faron, « Multi-Level Visual Tours of Weather Linked Data », in *Proceedings of VOILA'2022 co-located with the 21st International Semantic Web Conference (ISWC)* (Hangzhou, China), 2022.

[38] N. Yacoubi Ayadi, C. Faron, F. Michel, R. Bossy & A. Barbe, « Construction d'un graphe de connaissances à partir des annotations d'articles scientifiques et de leur contenu en sciences de la vie », in *IC'2022 – PFIA 2022 Journées francophones d'Ingénierie des Connaissances* (Saint-Etienne, France), 2022.

Abstract. — In this article, we present a knowledge graph built from a corpus of scientific articles on genomic selection methods for wheat culture. These main purpose of these methods is to improve the agronomic profile and quality of wheat varieties. The scientific literature on the subject has been growing steadily over the last twenty years. Initially, an NLP tool enabled us to extract and normalize various named entities by linking them to concepts previously defined in relevant domain ontologies. These entities refer to the names of genes, traits, phenotypes, markers and varieties (cultivars). The graph presented in this work structures and integrates these entities, based on the W3C Web Annotation Ontology (OA). The use of the OA ontology enables us to formalize the description of the context in which entities appear in the text. In this way, the graph highlights the context of appearance of these entities within the corpus, providing indications of the links and frequent associations between them. Based on a set of competency questions formulated by a domain expert, we validated the relevance of the proposed model and consequently the knowledge graph generated. In order to make our graph accessible to a large number of users, we developed several search and visualization interfaces enabling exploration of the contexts of appearance of several entities (of different types) in the same article. This work contributes to the structuring, understanding and exploration of knowledge in the field of wheat genomic selection, by providing a formal framework for the discovery and analysis of relationships between relevant entities in the domain. We propose an end-to-end knowledge engineering methodology that is both generic and adaptable, designed to facilitate the exploration and the analysis of scientific literature corpora across diverse disciplinary domains.

Keywords. — Linked Data, ontologies, semantic annotation, knowledge Graphs, text Mining.
