Construction d'un graphe de connaissances à partir des annotations d'articles scientifiques et de leur contenu en sciences de la vie

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

Construction d'un graphe de connaissances à partir des annotations d'articles scientifiques et de leur contenu en sciences de la vie

N. Yacoubi Ayadi[1], C. Faron[1], F. Michel[1], R. Bossy[2], A. Barbe[1]

[1] University Côte d'Azur, Inria, CNRS, I3S (UMR 7271), France
[2] MaIAGE, INRAE, Université Paris-Saclay, 78350 Jouy-en-Josas, France

yacoubi@i3s.unice.fr, faron@i3s.unice.fr, fmichel@i3s.unice.fr, robert.bossy@inrae.fr, arnaud.barbe@etu.univ-cotedazur.fr

Résumé

Dans ce papier, nous présentons un graphe de connaissances RDF permettant de décrire, structurer et intégrer des annotations d'entités nommées extraites automatiquement par l'outil Alvis NLP à partir de publications scientifiques portant sur la génétique et le phénotypage de blé. Ces entités nommées se réfèrent à la fois à des noms de gènes, traits, phénotypes, marqueurs et variétés impliqués dans la culture du blé. Cependant, une fois extraites, ces annotations sont stockées dans un format brut rendant difficile leur exploitation par les chercheurs. D'où, notre intérêt de les transformer (lifter) en un format compatible avec les standards de publication de données liées afin de construire un graphe de connaissances dans lequel des entités provenant à la fois de bases de connaissances génomiques et d'articles scientifiques ont été sémantiquement décrites et intégrées. Basé 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.

Mots-clés

Ontologie, Données liées, Annotation sémantique, Graphe de connaissances RDF, Fouille de textes.

Abstract

In this paper, we present an RDF knowledge graph to describe, structure and integrate annotations of named entities automatically extracted by the Alvis NLP tool from scientific publications on wheat genetics and phenotyping. These named entities refer to the names of genes, traits, phenotypes, markers and varieties involved in wheat breeding. However, once extracted, these annotations are stored in a raw format making it difficult for researchers to exploit them. Hence, our interest in transforming (lifting) them into a format compatible with linked data publication standards in order to build a knowledge graph in which knowledge coming from both genomic knowledge bases and scientific articles has been semantically described and integrated. Based on a set of competency questions formulated by a domain expert, we validated the relevance of the proposed model and consequently the generated knowledge graph.

Keywords

Ontologies, Linked Data, Semantic annotation, Knowledge Graph, Text Mining.

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 semanciers et 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. Ces techniques 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 (plus de 4000 sont publiées par an). Par conséquent, les techniques de TAL (Traitement Automatique de Langues naturelles) ont été largement utilisées pour la fouille de textes dans l'objectif d'extraire et de synthétiser les informations pertinentes pouvant aider les chercheurs dans leurs investigations.

Dans ce contexte, la plate-forme AlvisNLP [2] offre différents outils TAL pour l'extraction d'entités nommées à partir des publications scientifiques permettant ainsi d'annoter différents types d'entités à savoir des gènes, des phénotypes, des traits, des variétés, des marqueurs génétiques et des taxons. Toutefois, les résultats générés par cet outil sont exportés dans un format brut (i.e., CSV) ; ce qui entrave leur exploitation et intégration avec d'autres sources de connaissances.

Nous présentons, dans ce papier, le travail de recherche que

nous avons 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 complètement avec l'objectif du projet D2KAB et vise la construction d'un graphe de connaissances RDF intégrant des entités provenant de différents sources et processus, à la fois des bases de connaissances génomiques et des workflows TAL appliqués à la littérature scientifique. Ainsi, ce graphe de connaissances permettra d'intégrer les annotations extraites à partir des articles scientifiques avec d'autres ressources terminologiques et sémantiques publiées dans le cadre du Web sémantique. 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 [10] pour lifter les annotations produites par l'outil AlvisNLP en RDF, (2) l'utilisation d'un micro-service SPARQL pour récupérer les méta-données descriptives des publications scientifiques à partir de l'API PMC Entrez [2] et de les intégrer avec les annotations liftées. 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 validées par les experts.

Ce papier 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.

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) [13], BIBO (BIBliographic Ontology) et le vocabulaire Web Annotation Vocabulary (OA) [14].

2.1 Construction de graphes de connaissances à partir de textes

La problématique de construction de graphes de connaissances à partir de textes a suscité l'intérêt de plusieurs communautés incluant celles du Web sémantique, des données liées et du TAL. En effet, les techniques développées dans chacun de ces domaines s'avèrent complémentaires [9]. Une approche de construction de graphes à partir de textes doit combiner plusieurs outils et techniques pour permettre : (1) l'extraction et la reconnaissance d'entités nommées [5, 12], (2) le liage des entités nommées (normalisation d'entités nommées) à des concepts existants dans des ontologies/vocabulaires du domaine, (3) l'extraction de relations [12], et (4) la génération automatique du graphe RDF (RDFisation) [1] et sa publication conformément aux des principes des données liées. Le principal défi réside dans le liage des entités nommées. En effet, dans le domaine biomédical, les vocabulaires sont souvent volumineux et complexes. De plus, on observe un décalage important entre les étiquettes de concepts et les mentions dans le texte avec notamment l'usage extensif d'abréviations, de métonymies et de variations syntaxiques ([8], [3]). Le travail de recherche présenté dans [6] décrivant le challenge BioNLP-ST 2013 a mis en évidence l'intérêt de construire des graphes de connaissances RDF. Pour ce challenge, 10 bases de connaissances RDF ont été construites et évaluées à partir des annotations extraites par 10 systèmes TAL. L'objectif était de proposer un nouvel axe pour l'évaluation de la pertinence des annotations générées par ces systèmes. Ainsi, plusieurs requêtes SPARQL ont été conçues à l'aide d'experts du domaine et évaluées en comparant leurs résultats avec ceux obtenus de la base de connaissances de référence construite partir du *gold standard*. Contrairement aux auteurs dans [6] qui ont adopté un vocabulaire minimal conçu pour le besoin du challenge, nous nous basons sur l'utilisation conjointe de vocabulaires standards (i.e, [14]) et d'ontologies (i.e., [11]) pour la modélisation RDF des annotations générées a partir des publications scientifiques.

2.2 Vocabulaires et Ontologies existantes

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) [13] et BIBO [3] pour représenter les méta-données descriptives et bibliographiques des publications scientifiques. L'ontologie FaBio est une ontologie dérivée du modèle FRBR [4] 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 es-

1. http://www.d2kab.org/
2. https://www.ncbi.nlm.nih.gov/pmc/tools/developers/
3. https://github.com/structureddynamics/Bibliographic-Ontology-BIBO

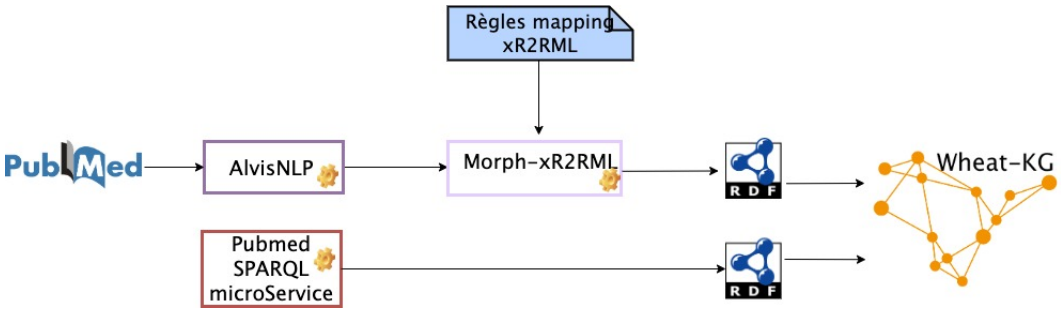

FIGURE 1 – Schéma général du pipeline de l'approche de construction du graphe Wheat-KG

sentiellement 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 [4] ou schema.org [5].

Enfin, le vocabulaire Web Annotation Vocabulary (OA) [14] 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 [7].

3 Questions de Compétences

Les questions de compétences permettent de résumer les exigences et les attentes des experts vis-a-vis d'un futur modèle de connaissances. Ces questions expriment les besoins des experts à explorer la littérature scientifique de la génomique de blé en exploitant les relations existantes entre les entités reconnues et annotées par AlvisNLP. Dans le contexte de ce travail de recherche, les CQ ont permis d'éliciter les attentes des chercheurs travaillant sur la génomique de blé et désirant exploiter la littérature scientifique autour de ce sujet. Ainsi, les CQ que nous présentons s'articulent autour des besoins des experts à explorer des possibles interactions entre les entités nommées en exploitant leur contexte de co-occurence dans le texte.

CQ1. Quelles publications scientifiques du corpus mentionnent le gène '*Lr34*' ?

CQ2. Quels sont les gènes mentionnés à proximité du phénotype *'drought tolerance'* ? Cette requête permet aux scientifiques de rechercher des gènes impliqués dans le contrôle d'un phénotype en particulier. Un ensemble de publications mentionnant un ou plusieurs gènes apparaissant avec le phénotype en question sont retournées comme résultat.

CQ3. Quels sont les marqueurs génétiques mentionnés à proximité d'un gène, qui lui-même est mentionné à proximité d'un phénotype particulier ? Cette requête permet de rechercher des publications mentionnant des marqueurs qui pourraient servir à sélectionner un phénotype donné. Comme les techniques de marquage génétique ont évolué au fil du temps et certaines sont devenues obsolètes, l'expert peut également raffiner sa requête pour sélectionner uniquement les publications apparues après 2010. D'où l'intérêt de représenter dans le graphe des méta-données descriptives telles que la date d'apparition de la publication.

CQ4. Quelles sont les variétés de blé qui présentent un phénotype particulier ? L'expert peut rechercher des variétés d'intérêt car elles présentent un phénotype spécifique.

CQ5. Effectuer une recherche bibliographique de toutes les publications mentionnant des gènes spécifiques à des variétés de blé tendre (*Triticum aestivum*) qui présentent un phénotype général. Le résultat à cette requête devra inclure une liste d'articles mentionnant à la fois des gènes, une ou plusieurs variétés de blé tendre et le phénotype en question. Cependant, si l'expert est intéressé par les gènes impliqués dans la résistance aux pathogènes, il faudrait alors inclure dans les résultats de la requête tout type de pathogènes (bactérie, virus, champignons). D'où l'intérêt d'intégrer dans le graphe des connaissances ontologiques et terminologiques qui proviennent des ontologies et des vocabulaires du domaine.

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.

4.1 Description des articles scientifiques

Pour représenter et décrire les publications scientifiques, nous avons réutilisé les vocabulaires suivants : Dublin Core, FRBR aligned Bibliographic Ontology (FaBiO) et Bibliographic Ontology (BIBO). 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

4. https://www.dublincore.org/specifications/dublin-core/dcmi-terms/
5. https://schema.org/

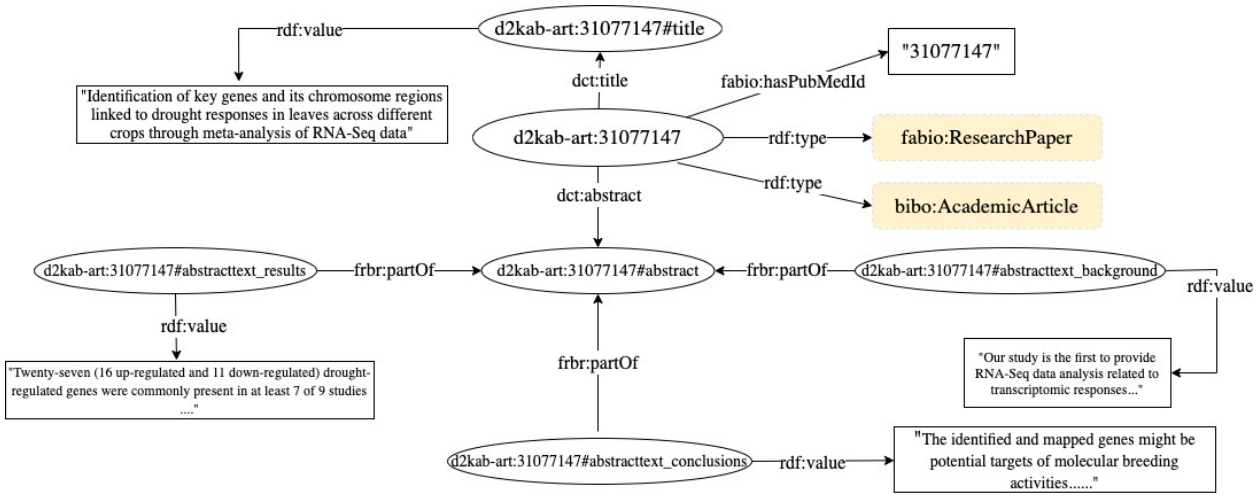

FIGURE 2 – Exemple de graphe RDF représentant une publication du corpus

lien *partie de* entre un résumé et une publication et aussi entre un résumé et ses sous-sections. La figure 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*).

4.2 Description des annotations extraites des articles scientifiques

Une annotation A_i est une indication qu'une mention m_e d'une entité nommée e a été identifiée dans le résumé (ou l'une de ses sous-sections) d'un article a à une position de début d et une position de fin f. Dans ce travail, nous réutilisons le vocabulaire OA pour représenter les annotations d'entités nommées. Ainsi, une annotation A_i est représentée comme une instance de la classe `oa:Annotation` et est décrite par les informations suivantes :

- A_i a une cible représentée comme l'objet de la propriété `oa:hasTarget`. Cette cible représente une occurrence d'une entité nommée e identifiée par la présence de sa mention m_e (*surface form*) entre une position de début d et une position de fin f dans le texte du résumé ou une de ses sous-sections.
- A_i a un corps `oa:hasBody` qui renvoie vers l'URI d'une entité e définie dans une ontologie ou un vocabulaire du domaine tels par exemple l'URI d'un concept WTO [11] ou d'une classe de taxons dans la taxonomie NCBI [6].

La figure 3 présente les différentes classes d'entités de notre graphe de connaissances. Comme le montre cette figure, ces classes sont inter-connectées par des relations sémantiques.

La figure 4 illustre un exemple d'annotation extraite à partir de la sous-section conclusion d'un résumé identifiée par l'URI `d2kab-art:31077147#abstracttext_conclusions`. Dans cette sous-section, la mention du trait '*drought resistance*' a été identifiée et mise en correspondance avec l'entité WTO `wto:WTO_0000311`. La position de début et de fin, ainsi que la valeur de la mention m_e sont indiquées respectivement par l'utilisation des propriétés : `oa:start`, `oa:end` et `oa:exact`. Cette représentation structurée en RDF permet de modéliser à la fois les portions des publications à partir desquelles une annotation a été extraite et d'autre part leur mise en correspondance avec des entités prédéfinies dans des ontologies et des vocabulaires du domaine. Ceci offrira la possibilité aux chercheurs d'explorer les contextes d'occurrence et de co-occurrence des entités dans les textes scientifiques.

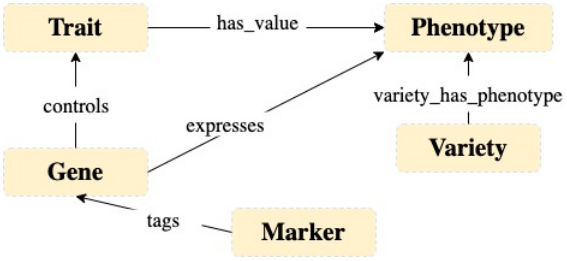

FIGURE 3 – Les classes d'entités extraites Alvis NLP et leurs relations sémantiques

5 Construction du graphe de connaissances

Nous présentons dans cette section le jeu de données et les outils utilisés pour la construction du graphe de connaissances. La figure 1 illustre le processus de construction graphe Wheat-KG à partir de la littérature scientifique de la génomique des plantes représentée par un corpus d'articles extraits à partir de PubMed [7].

5.1 Jeu de données

Le jeu de données fourni par l'équipe MaIAGE comprend plusieurs types d'informations stockées séparément dans

6. NCBI Taxonomy https://www.ncbi.nlm.nih.gov/taxonomy

7. https://pubmed.ncbi.nlm.nih.gov/

	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 1 – Template de génération d'URI des ressources de notre graphe

Classe	
Nbre total d'annotations	88880
Nbre total d'articles	8496
Nbre total de gènes	1160
Nbre total de taxons	2462
Nbre total de phénotypes	98
Nbre total de marqueurs	521
Nbre total de variétés	77
Nbre total de relations	162

TABLE 2 – Nombre d'entités par classes dans notre graphe de connaissances

plusieurs fichiers CSV. Ainsi, ce jeu inclut un corpus constitué de 8496 publications à partir desquels un ensemble de 4318 entités nommées a été extrait en utilisant la plateforme AlvisNLP qui offre une chaîne de traitement TAL pour l'annotation sémantique de documents textuels dans le domaine de la génomique des plantes. Cette plate-forme 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. 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, 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.

5.2 Processus de lifting

Pour transformer les annotations produites par AlvisNLP en un graphe RDF, nous avons utilisé l'outil morph-xR2rml [10]. 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 [8]. 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 [9]. La table 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 avons utilisé un micro-service SPARQL [10] qui permet d'interroger l'API PubMed Central et de récupérer les méta-données 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 4.1. Le tableau 2 représente le nombre de triplets pour chaque classe du graphe de connaissances.

6 Validation

Plusieurs CQ définies par un expert du domaine exprimant des besoins d'explorer la littérature scientifique selon différents critères sont présentées dans la section 3. Toutes les CQ ont été traduites en requêtes SPARQL [11] et les résultats ont été validés par l'expert. Cependant, nous nous contentons de présenter dans cette section uniquement les requêtes SPARQL implémentant les questions de compétence *CQ4* et *CQ5*. Ces requêtes montrent comment notre graphe de connaissances peut être exploité pour répondre aux besoins d'explorer la littérature discutant des gènes, des phénotypes et variétés de blé.

CQ4. Quelles sont les variétés de blé qui présentent un phénotype particulier ? L'intention de cette requête est de rechercher uniquement des variétés d'intérêt présentant le phénotype spécifié. En effet, dans le graphe de connaissances, les variétés sont inter-reliées à des phénotypes par la relation "`variety_has_phenotype`". le listing 1 présente la requête SPARQL correspondante à la CQ4.

```
SELECT distinct ?variety  ?document
WHERE {
?relation1 d2kab:hasVariety ?aVariety ;
           d2kab:hasPhenotype ?aPhenotype .
?aVariety a oa:Annotation ;
           oa:hasTarget ?t1 ;
           oa:hasBody ?Variety .
?t1 oa:hasSource ?partDoc1 .
?Variety a d2kab:Variety ;
           skos:prefLabel ?variety.
?aPhenotype a oa:Annotation ;
           oa:hasTarget ?t2 ;
           oa:hasBody ?Pheno .
?t2 oa:hasSource ?partDoc2 .
?Pheno skos:prefLabel 'drought tolerance' .
?partDoc1 frbr:partOf+ ?document .
?partDoc2 frbr:partOf+ ?document .
?document  a fabio:ResearchPaper .
}
```

Listing 1 – Requête SPARQL de la CQ4

8. https://www.i3s.unice.fr/~fmichel/xr2rml_specification_v5.html
9. https://github.com/Wimmics/d2kab-wheat-kg/tree/main/Mapping_rules
10. https://sparql-micro-services.org/service/pubmed/getArticleByPMId_sd/
11. https://github.com/Wimmics/d2kab-wheat-kg/tree/main/sparql-queries

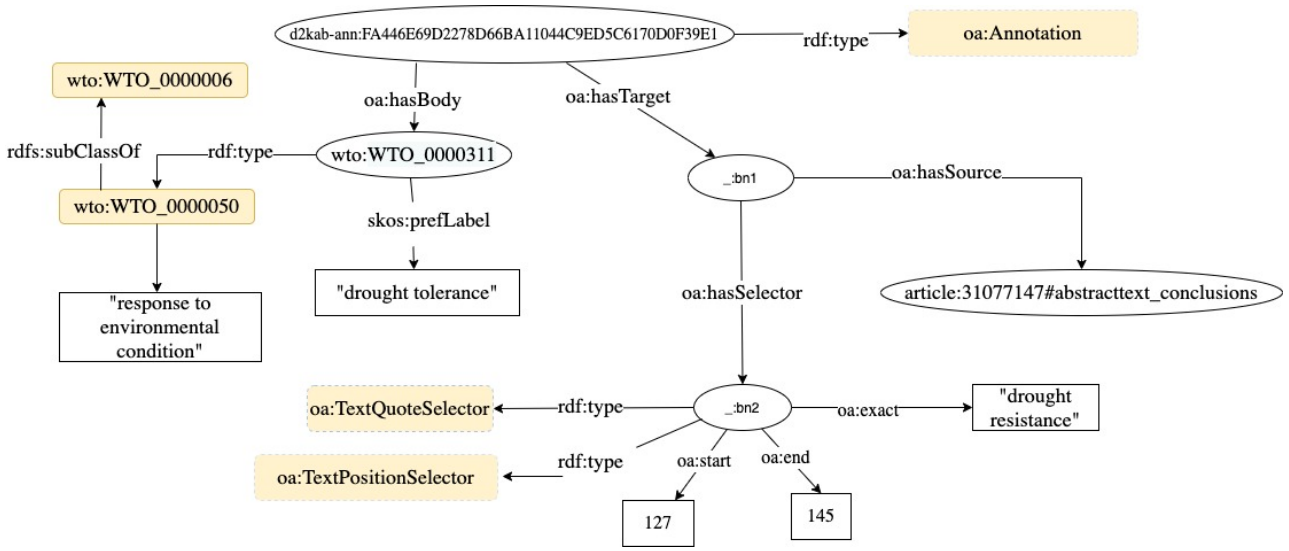

FIGURE 4 – Exemple de graphe RDF modélisant une annotation de l'entité *drought resistance* présente de la position 127 à 145 dans la sous-section conclusion du résumé d'un article

CQ5. Effectuer une recherche bibliographique de toutes les publications mentionnant des gènes spécifiques à des variétés de blé tendre (Triticum aestivum) qui présentent un phénotype général. Le résultat de cette requête devra inclure une liste de publications ayant mentionnées à la fois des gènes, une variété de blé tendre et un phénotype particulier. Si l'expert est intéressé par les gènes impliqués dans l'expression de phénotypes de résistance aux pathogènes, il faudrait alors inclure dans le résultat de la requête les phénotypes de résistance à tous les sous-types de pathogènes (bactérie, virus, champignons) qui peuvent obtenus a partir du thésaurus WTO. En effet, la requête du listing 2 exploite les relations hiérarchiques entre les concepts WTO pour inclure toutes les publications mentionnant des sous-concepts du phénotype '*pathogen resistance*'.

```
SELECT distinct ?doc ?gene ?variety
WHERE {
?rel1 d2kab:hasVariety ?aVariety ;
    d2kab:hasPhenotype ?aPhenotype.

?aVariety a oa:Annotation ;
            oa:hasTarget ?t1 ;
            oa:hasBody ?Variety .
?t1 oa:hasSource ?d1 .
?Variety a d2kab:Variety ;
        skos:prefLabel ?variety.
?aPhenotype a oa:Annotation ;
            oa:hasTarget ?t2 ;
            oa:hasBody ?Phenotype .
?t2 oa:hasSource ?d2 .
?e2 skos:prefLabel "pathogen resistance" ;
    skos:narrower* ?Phenotype .
?aTaxon a oa:Annotation ;
        oa:hasTarget ?t ;
        oa:hasBody "Triticum aestivum".
?t oa:hasSource ?d .
?Taxon a d2kab:Taxon ;
        skos:prefLabel ?taxon.
?aGene a oa:Annotation ;
        oa:hasTarget ?t3;
        oa:hasBody ?Gene.
?t3 oa:hasSource ?d3 .
?Gene a d2kab:Gene ;
        skos:prefLabel ?gene.
?d1 frbr:partOf+ ?doc .
?d2 frbr:partOf+ ?doc .
?d3 frbr:partOf+ ?doc .
?d frbr:partOf+ ?doc .
?doc  a fabio:ResearchPaper .
}
```

Listing 2 – Requête SPARQL de la CQ5

7 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 Wheat-KG 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 Wheat-KG, 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, [14]) et des ontologies du domaine en question (i.e., [11]). 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 en-

visageons d'intégrer dans le graphe de connaissances des observations collectées par des professionnels du domaine. Ces observations décrivent les stades de croissances 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.

8 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)

Références

[1] Alberto Anguita, Miguel Garcia-Remesal, Diana de la Iglesia, and Víctor Maojo. NCBI2RDF : enabling full RDF-based access to NCBI databases. *BioMed research international*, 2013 :983805, 01 2013.

[2] Mouhamadou Ba and Robert 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*, pages 15–18, Portorož, Slovenia, 2016.

[3] Robert Bossy, Louise Deleger, Estelle Chaix, Mouhamadou Ba, and Claire Nédellec. Bacteria Biotope at BioNLP Open Shared Tasks 2019. In *5th Workshop on BioNLP Open Shared Tasks BioNLP-OST@EMNLP-IJCNLP 2019, Association for Computational Linguistics*, page np, Hong-Kong, Hong Kong SAR China, November 2019. Jin-Dong Kim and Claire Nédellec and Robert Bossy and Louise Deléger.

[4] Ian Davis and Richard Newman. Expression of core FRBR concepts in RDF. https://vocab.org/frbr/core.

[5] John M Giorgi and Gary D Bader. Towards reliable named entity recognition in the biomedical domain. *Bioinformatics*, 36(1) :280–286, 06 2019.

[6] Jin-Dong Kim, Jung-Jae Kim, Xu Han, and Dietrich Rebholz-Schuhman. Extending the evaluation of genia event task toward knowledge base construction and comparison to gene regulation ontology task. *BMC bioinformatics*, 16 :S3, 07 2015.

[7] Jin-Dong Kima, Karin Verspoorb, Michel Dumontierc, and K Bretonnel Cohend. Semantic representation of annotation involving texts and linked data resources. *Semantic Web journal*, 2015.

[8] Robert Leaman, Ritu Khare, and Zhiyong Lu. Challenges in clinical natural language processing for automated disorder normalization. *Journal of biomedical informatics*, 57 :28–37, 2015.

[9] José-Lázaro Martínez-Rodríguez, Aidan Hogan, and I. Lopez-Arevalo. Information extraction meets the semantic web : A survey. *Semantic Web*, 11 :255–335, 2020.

[10] Franck Michel, Loïc Djimenou, Catherine Faron-Zucker, and Johan Montagnat. Translation of relational and non-relational databases into RDF with xR2RML. In Valérie Monfort, Karl-Heinz Krempels, Tim A. Majchrzak, and Ziga Turk, editors, *WEBIST 2015 - Proceedings of the 11th International Conference on Web Information Systems and Technologies, Lisbon, Portugal, 20-22 May, 2015*, pages 443–454. SciTePress, 2015.

[11] Claire Nédellec, Liliana Ibanescu, Robert Bossy, and Pierre Sourdille. WTO, an ontology for wheat traits and phenotypes in scientific publications. *Genomics & Informatics*, 18, 2020.

[12] Nadeesha Perera, Matthias Dehmer, and Frank Emmert-Streib. Named entity recognition and relation detection for biomedical information extraction. *Frontiers in Cell and Developmental Biology*, 8, 2020.

[13] Silvio Peroni and David Shotton. FaBiO and CiTO : Ontologies for describing bibliographic resources and citations. *Journal of Web Semantics*, 17 :33–43, 2012.

[14] Robert Sanderson, Paolo Ciccarese, and Benjamin Young. Web annotation ontology. https://www.w3.org/TR/annotation-vocab/, 2017.