Smart Agricultural Technology 8 (2024) 100484

Contents lists available at ScienceDirect

Smart Agricultural Technology

journal homepage: www.journals.elsevier.com/smart-agricultural-technology

A unified approach to publish semantic annotations of agricultural documents as knowledge graphs

Nadia Yacoubi Ayadi [a,b,*], Stephan Bernard [c,g], Robert Bossy [d], Marine Courtin [c,d], Bill Gates Happi Happi [e], Pierre Larmande [e], Franck Michel [b], Claire Nédellec [d], Catherine Roussey [c,f], Catherine Faron [b]

[a] Université Claude Bernard Lyon 1, CNRS, LIRIS (UMR 5205), France
[b] Université Côte d'Azur, Inria, CNRS, I3S (UMR 7271), France
[c] TSCF, INRAE, Centre Auvergne Rhône Alpes Clermont, Aubière, France
[d] MaIAGE, INRAE, Université Paris-Saclay, 78350 Jouy-en-Josas, France
[e] DIADE, IRD, CIRAD, Univ. Montpellier, Montpellier, France
[f] Mistea, INRAE, Centre Occitanie, Montpellier, France
[g] LISC, INRAE, Centre Auvergne Rhône Alpes Clermont, Aubière, France

ARTICLE INFO

Keywords:
Agriculture
Knowledge graphs
Semantic modelling
RDF transformation
Natural language processing
Annotations
Semantic resources
Named entity recognition and linking

ABSTRACT

The research results presented in this paper were obtained as part of the D2KAB project (Data to Knowledge in Agriculture and Biodiversity) which aims to develop semantic web-based tools to describe and make agronomical data actionable and accessible following the FAIR principles. We focus on constructing domain-specific Knowledge Graphs (KGs) from textual data sources, using Natural Language Processing (NLP) techniques to extract and structure relevant entities. Our approach is based on the formalization of a semantic data model using common linked open vocabularies such as the Web Annotation Ontology (OA) and the Provenance Ontology (PROV). The model was developed by formulating motivating scenarios and competency questions from domain experts. This model has been used to construct three different KGs from three distinct corpora: PubMed scientific publications on wheat and rice genetics and phenotyping, and French agricultural alert bulletins. The named entities to be recognized include genes, phenotypes, traits, genetic markers, taxa and phenological stages normalized using semantic resources such as the Wheat Trait and Phenotype Ontology (WTO), the French Crop Usage (FCU) thesaurus and the Plant Phenological Description Ontology (PPDO). Named entities were extracted using different NLP approaches and tools. The relevance of the semantic model was validated by implementing experts questions as SPARQL queries to be answered on the constructed RDF knowledge graphs. Our work demonstrates how domain-specific vocabularies and systematic querying of KGs can reveal hidden interactions and support agronomists in navigating vast amounts of data. The resources and transformation pipelines developed are publicly available in Git repositories.

1. Introduction

Knowledge Graphs (KG) are multi-relational graphs of relations between well-defined and uniquely identifiable entities created from heterogeneous data sources. They enable to develop data management platforms compliant with the FAIR (Findability, Accessibility, Interoperability and Reuse) principles [27] referring to best practice guidelines: resources must be accessible, understood, exchanged and reused by machines. In this context, the goal of the D2KAB[1] project (Data to Knowledge in Agriculture and Biodiversity), of which this work is a part, is to develop new semantic web-based tools for the semantic description of agronomical and agricultural data, making them actionable, and openly accessible, according to the FAIR principles. A typical approach towards publishing FAIR knowledge graphs is to rely on Linked Data (LD) principles and Semantic Web technologies (SWT). Indeed, RDF and other Semantic Web standards are designed to promote interoperability and

* Corresponding author.
E-mail address: nadia.yacoubi-ayadi@univ-lyon1.fr (N. Yacoubi Ayadi).
[1] https://d2kab.mystrikingly.com/.

https://doi.org/10.1016/j.atech.2024.100484
Received 15 March 2024; Received in revised form 4 June 2024; Accepted 4 June 2024
Available online 6 July 2024

linking between datasets. Additionally, to ensure that RDF datasets are truly interoperable and reusable within a specific field, they must rely on domain-specific and open vocabularies, models, and data category registries capturing the shared theoretical foundations and terminology used by a community of agronomists and farmers [6,4]. Constructing knowledge graphs from unstructured data enables to bridge the gap between the huge amount of heterogeneous data and easily explore and query it to address various use cases. This paper focuses on building knowledge graphs from textual data sources by extracting relevant domain-specific entities and organizing them in structured and meaningful annotations. This approach can be beneficial in making sense of large, complex and heterogeneous datasets, linking related information and knowledge, and providing intuitive ways to access and explore domain data and knowledge, adhering to the FAIR principles. In this paper, we present a methodology for constructing domain-specific knowledge graphs using SWT, which involves the re-use of shared RDF-based vocabularies and models. Although the proposed methodology can be applied to various domains and can support a wide range of use cases, in this research we focus on building knowledge graphs representing semantic annotations of textual documents in the field of agriculture and agronomy. We consider three different text corpora and we demonstrate how we leverage Natural Language Processing (NLP) techniques to first extract different types of named entities and then structure and integrate them into KGs using the same data model. Thus, we demonstrate how it is possible to jointly query different knowledge graphs built from different text corpora in order to enable knowledge discovery. Two of the corpora are collections of scientific publications on rice and wheat functional genomics respectively, retrieved from the PubMed[2] repository. These publications investigate the gene-phenotype link for varietal selection, and more precisely the identification of gene markers involved in the expression of a given phenotype, for selection assistance [15]. Despite containing valuable descriptions of phenotypes linked to cultivars and genetic data, the scientific literature remains largely underutilized [16]. Identifying and understanding repeated co-occurrences of genes, traits, and phenotypes are crucial for discovering possible interactions between them and accordingly selecting and improving wheat or rice varieties with desirable characteristics (increased diseases or stress resistance, for instance). A third corpus gathers technical documents called Plant Health Bulletins (PHBs) which are agricultural alert bulletins published in France. These bulletins are dispersed across regional websites,[3] making it challenging to query the entire corpus and retrieve results by crop. Moreover, we would like to retrieve alert bulletins based on spatiotemporal thematic annotations and computes related aggregation indicators as presented in [20]. Thus, KGs built from these corpora are intended to meet the following needs:

- Navigating through an exponential number of documents can be challenging for agronomists. Extracting and annotating relevant entities using domain-specific vocabularies can greatly help them to navigate efficiently through vast corpora. This research demonstrates how it is possible to uniformly query corpora which have different focus; scientific literature from PubMed and in-field observational data described in PHB documents.
- Cross-referencing entities to identify network of frequently co-occurring entities within documents can reveal potential interactions between them.
- Enabling semantic search capabilities allows agronomists and farmers to quickly retrieve specific information and relationships from large volumes of unstructured scientific text and in-field observations, enhancing the efficiency of knowledge discovery and hypothesis generation.

The first step of our methodology leverages NLP pipelines to perform the tasks of Named Entity Recognition (NER) and Linking (NEL) [14]. A semantic data model has been defined to capture how NE annotations produced by NLP pipelines should be structured and described in each KG. We initiated the data model definition with a set of competency questions, together with motivating examples. We were inspired by the agile SAMOD methodology [17], which in turn is based on the early work of Uschold & Gruninger [25]. The SAMOD process is initiated by a motivating scenario which lead to a set of competency questions (CQs) that provide requirements on the knowledge graphs to be created. CQs have highlighted the need for consistent and systematic querying of KGs by agronomists in order to reveal hidden interactions between NE that co-occur within the same context in scientific documents. Thereby, several CQs have stressed the importance of identifying frequently associated entities, such as genes or varieties and phenotypic traits in functional genomics corpus or cultivated crops in French regions in the PHB corpus.

Inline with earlier works [12], we propose to rely on the Open Annotation Ontology (OA) [23] to describe, structure and integrate NE annotations and their occurrence contexts in texts. Domain specific vocabularies are also reused to describe bibliographic information of PubMed publications and provenance information for the PHB corpus. The resulting model was automatically populated using a mapping-based transformation pipeline implemented with the Morph-xR2RML tool [8].

The text corpora we collected, the Knowledge Graphs we built, the data transformation pipelines with the mapping rules we built are available and documented in Git repositories.[4,5,6]

The paper is structured as follows. In Section 2, we present the materials of our research work consisting of three text corpora and the semantic resources used to annotate them. Section 3.1 outline the competency questions (CQs) of each case study. We provide a detailed description of the proposed semantic model in Section 3.2. Section 4 presents the validation results of the case studies. Finally, in Section 5, we discuss the results and synthesize the learned lessons before concluding in Section 6.

2. Materials

In this section, we present the different materials used in this research work to build our KGs. In Section 2.1, we first present the text corpora that were processed using different NLP pipelines in order to generate semantic annotations of domain named entities. The Wheat and PHB corpora were processed using the AlvisNLP corpus processing engine [1] that allows to assemble pipelines according to specific needs. Indeed, these two corpora are written in distinct languages (English and French respectively), contain documents of distinct genre (scientific and occupational), and require the extraction of distinct named entity types. Each customized pipeline relied on unsupervised methods that exploit specific lexico-semantic resources, such as the WTO or the BBCH-based Plant Phenological Description Ontology (semantic resources presented in Section 2.2). In contrast, the DIADE group employed the HunFLAIR NER tagger in the rice genomics corpus as discussed in section 2.1.2.

2.1. Text corpora

2.1.1. PubMed corpus on wheat genomics

The Wheat corpus was collected by the MaIAGE research group.[7] The corpus consists in 8,496 scientific references of articles related to

[2] https://pubmed.ncbi.nlm.nih.gov/.

[3] See for example the web page of the French Minister of Agriculture https://agriculture.gouv.fr/bulletins-de-sante-du-vegetal.

[4] https://github.com/Wimmics/WheatGenomicsSLKG.

[5] https://github.com/ANR-DIG-AI/RiceGenomicsSLKG.

[6] https://forgemia.inra.fr/bsv/corpus-bsv.

[7] https://maiage.inrae.fr/en.

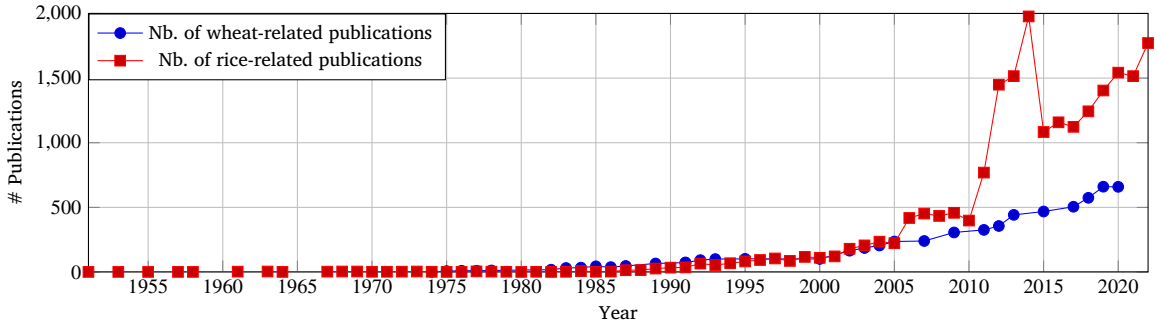


Fig. 1. Evolution of the number of research works on wheat and rice genomics from 1951 to 2022.

wheat selection published between 1974 and 2021 which covers the early stages of the field to the current time of this work. A first exploration of the corpus shows that over the last two decades, the number of publications on has steadily increased. More than 80% of the publications in the corpus originate from the last two decades, which reflects the significant rise of research interest in wheat genomics as shown in Fig. 1.

In this corpus, the PubMed identifier, title and abstract of each publication are provided. In several cases, the abstract of a publication is organised in different sub-sections (i.e., background, results and conclusion). The AlvisNLP pipeline focused on the information relevant to genetic marker-assisted selection. These entities include genes, traits, phenotypes, taxa and varieties mentioned in the title and the abstract of publications, as well as the relationships between wheat varieties and phenotypes. In total, 88,880 mentions of 4,318 distinct named entities were recognized and linked to existing entities of semantic resources (presented in Section 2.2). Fig. 4 illustrates an example of PubMed publication where three types of NEs are recognised: we distinguish between NE mentions that refer to genes (e.g., *Sr2*, *Lr27*, *Lr34*), traits (e.g., *leaf rust resistance*, *resistance to stem rust*, *powdery mildew resistance*) and taxa (e.g., *wheat*).

The trait and phenotype mentions are linked to classes and concepts in the Wheat Trait Ontology (WTO) and taxon mentions are linked to NCBI taxonomy classes.

2.1.2. PubMed corpus on rice genomics

The DIADE research group[8] collected 17,058 scientific articles from the Oryzabase database [7] which provides manually checked PubMed entries related to rice genomics. The corpus represents scientific articles published between 1951 and 2021. It is worth noticing an increasing pace of publishing activity during this period, first coinciding with the availability of the two main rice genomes and their annotations (2004-2008) and then to the development of second and third generation of sequencing techniques (2012-2018) allowing faster and cheaper genome sequences availability. Most of the articles in the corpus date from the last two decades, which is consistent with the sharp increase in research interest in rice genetics as depicted in Fig. 1.

We used the HunFLAIR NER tagger [28] known for its superior performance across a variety of biomedical datasets, leveraging embeddings and pre-trained models specifically tailored for biomedical literature. We combined it with sciSpaCy using its specialised tokenizer for biomedical NER and other Python libraries to extract four types of named entities in the title or the abstract of the articles. The choice of HunFLAIR was driven by its proven efficiency in accurately identifying and classifying biological entities (genes, proteins, species, diseases, phenotypes), which are crucial for our analysis to identify potential agronomic traits of interest associated with molecular entities. We distinguished between NE mentions that refer to genes (e.g., *OsMAPK2* or *MOC1*), species (e.g., *Oryza sativa* or *Magnaporthe oryzae*), chemicals (e.g., *gibberellic acid* or *nitrogen*) and diseases or phenotypes (e.g., diseases *blast* or *Sheath blight disease*). In total, 351,003 mentions of 63,591 distinct NEs were identified from PubMed abstracts and titles. When possible, these NEs were linked with existing semantic resources as explained in Section 2.2.

For both PubMed corpora, we considered the titles and abstracts of scientific papers, not the full text. Processing the full texts would allow the extraction of significantly much more information but would be much more time-consuming and therefore challenging. Moreover, full-texts are not always legally or technically available for processing; in this work we focused on thoroughness and openness.

2.1.3. Plant health bulletin corpus

In France, the Grenelle Environment and Ecophyto 2018 program strengthened national surveillance networks of crops and agricultural practices. Plant Health Bulletins are one of the modalities established by these surveillance networks in all regions and French overseas departments. A Plant Health Bulletin (PHB) is an agricultural alert document, both technical and regulatory in nature, written in French under the responsibility of a regional epidemiological surveillance committee. A PHB gathers information about the health status of crops. It reports observations of crop development and pest attacks, and analyses pest risk in the whole area. Nearly 15,000 plots are observed each year to edit approximately 3400 PHBs per year [20]. PHBs synthesize the interpretation of observations performed on crops by different collecting networks, elements from epidemiological models, meteorological data and sometimes biological analysis. Thus, the PHB corpus can be seen as a French archive of human validated crop observations on the whole French territory.

The TSCF research group (Technologies and Information Systems for Agrosystems) has collected 36,469 bulletins from 2009 to 2022 from the whole French territory. In this work, we considered three sub-corpora of PHBs previously used to validate NLP processes: the Vespa corpus gathers 497 PHBs collected in the whole French territory between 2009 and 2015; the D2KAB corpus is composed of 230 PHBs collected in 2019, manually selected to cover the whole French territory and to represent three crop categories – field crops, vegetables and grapevines; and the Alea corpus is composed of 150 PHBs randomly selected from the whole corpus. Overall, the publication date may vary from 2009 to 2020. These three sub-corpora are available on a Git repository.[9] The whole corpus brings together a total of 877 PHBs with an average of 2,548 tokens per bulletin, covering the whole French territory and all crop categories of French agriculture. These corpora were processed with a custom AlvisNLP pipeline, designed to extracted NEs referring to french crop names and french development stages mentioned in the text of PHBs. This specific pipeline is available in the git repository mentioned above. Thus, we would be able to study observations of crop development over

[8] http://diade.ird.fr/.

[9] https://forgemia.inra.fr/bsv/corpus-bsv.

Fig. 2. Example of expected NE recognition and linking in a grapevine PHB.

time according to different dimensions: per crop, per region and per year in order to characterize the impact of climate change on crops. In total, 72,993 mentions of 461 distinct NEs were extracted. Fig. 2 illustrates an example of PHB where two types of NEs are recognised. We distinguish NE mentions that refer to french crop names: e.g., *viticulture*, *fleurs* (flowers), *baies* (berries), *Melon*, *pois* (peas); and french development stages: e.g., *floraison* (flowering), *BBCH 69*, *BBCH-69 et 73*, *BBCH-75*, *développement des fruits* (fruit development). Those mentions are linked to existing elements defined in the FCU thesaurus and the BBCH-based Plant Phenological Description Ontology.

2.2. Semantic resources

In the agriculture domain, an increasing number of semantic resources (ontologies, thesauri) was developed and published using Semantic Web technologies [4] and made available for research communities in open portals such the Agroportal repository[10] [5]. In this section, we present the semantic resources that we have reused to annotate the text corpora presented in Section 2.1.

2.2.1. Wheat trait and phenotype ontology

The Wheat Trait and Phenotype Ontology (WTO) [16] is a domain ontology that covers a wide range of wheat traits and phenotypes related to soft wheat (*Triticum aestivum L.*) and the environmental factors that affect these traits. While traits denote physical observable plant properties, phenotypes are the set of possible values of traits. Capturing phenotypic information in a formal, shared representation is crucial for scientists as well as for breeders. However, automatic annotation of textual data remains a challenge due to the large number of traits and the great diversity of the vocabulary used to designate them. WTO has been developed to meet the requirements of trait and phenotype annotation in the scientific literature. In WTO, traits are organised into different categories such as development, morphology, quality, response to environmental conditions including biotic and abiotic stresses.

WTO (3.0) is available in the OBO format on Agroportal[11] and contains 745 classes. The transformation of WTO addresses the need for semantic integration in the Linked Open Data, facilitating experts to uniformly query KGs annotating scientific literature on phenotypic information linked to this reference vocabulary formalized in Semantic Web standards. We used WTO to annotate mentions of phenotypes and traits recognised in the PubMed corpus on wheat functional genomics.

2.2.2. NCBI taxonomy

The NCBITaxon ontology[12] is an automatic translation of the NCBI taxonomy into OWL. The NCBI Taxonomy consists of a single, hierarchically arranged list of organismal names across all domains of life. These names are correct, current and valid according to the best authorities within the separate taxonomic disciplines and codes of nomenclature [22]. In the NCBITaxon ontology, the NCBI taxons are translated into OWL classes whose instances would be individual organisms. The labels of NCBITaxon classes are the scientific names (e.g. *Triticum aestivum L*) and vernacular names (e.g. *soft wheat*) of the taxons. We used NCBITaxon to annotate and link different types of organisms mentioned in both PubMed corpus on wheat and rice functional genomics, including species, viruses and pathogens.

[10] http://agroportal.lirmm.fr/.

[11] http://agroportal.lirmm.fr/ontologies/WHEATPHENOTYPE?p=summary.

[12] https://obofoundry.org/ontology/ncbitaxon.html.

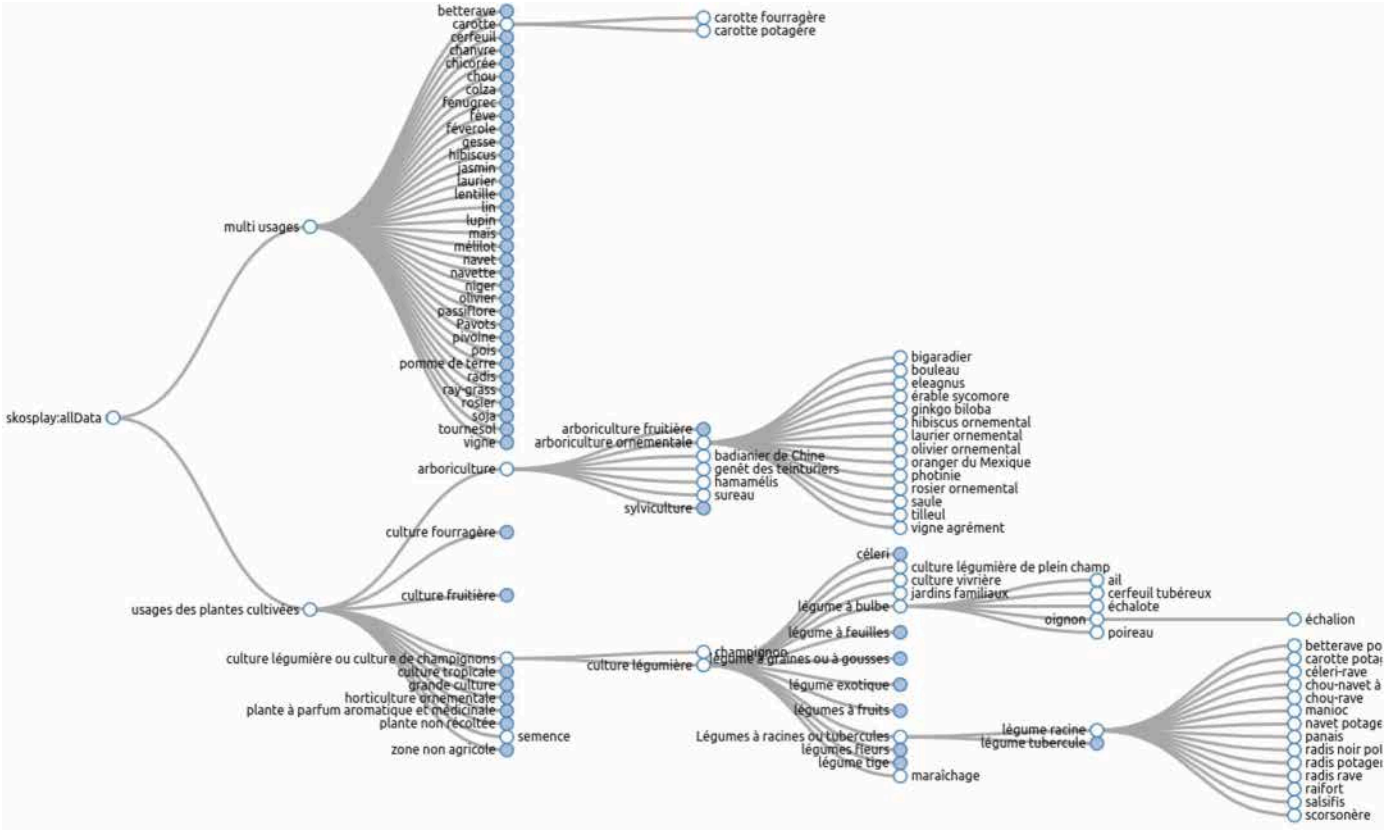


Fig. 3. An extract from the FCU thesaurus. Visualisation generated by the SKOS Play tool.

2.2.3. French crop usage thesaurus

The French Crop Usage (FCU) thesaurus[13] organises plants based on their roles in agriculture, or in other words, agricultural plant uses. The thesaurus hierarchy has two main branches as shown in Fig. 3. The branch named *Multiusages* contains all the cultivated plants that have several uses in agriculture. For example, *carotte* (carrot) may be used as vegetable or as fodder. The branch *Usages_plantes_cultivees* organises cultivated plants according to their uses and represents crop categories. FCU stores only the french vernacular names of plants. The FCU thesaurus is formalized using SKOS and used in this work to extract and link crop names in the PHB corpus.

2.2.4. BBCH-based plant phenological description ontology

BBCH (*Biologische Bundesanstalt, Bundessortenamt und CHemische Industrie*) is considered as a reference to describe development stages of different plant species in four languages: English, French, Spanish and German. It describes several sets of development stages. A set of stages composes a BBCH scale. Some plant species (like tomatoes or potatoes) have a specific set of stages named 'individual scales'. A general BBCH scale is also defined for plant species where no individual scale exists [9]. BBCH framework uniformly codes phenologically similar development stages of different plant species.

The BBCH-based Plant Phenological Description Ontology (PPDO)[14] [21] relies on the BBCH scale. It formalizes the scales and their associated stages as a specialization of the SKOS model. A stage is an instance of class *skos:Concept* with labels and definitions in four languages. The BBCH scale is represented as a SKOS thesaurus. The BBCH thesauri are used to extract and link phenological stages in the PHB corpus.

2.2.5. Resources for other mentions

Resources for genes, markers, wheat and rice varieties published as LOD datasets are very limited and in most cases, they are either incomplete, do not come from authoritative organizations, or do not provide unique identifiers. Among the available semantic resources for genes and markers, the UniProt Knowledge base [24] (UniProtKB) is a central hub for a collection of functional information on proteins, with accurate, consistent and rich annotation which is accessible through a SPARQL endpoint.[15] However, multiple UniProtKB identifiers can be retrieved for the same genomic entity which makes it impossible to link named entities using this resource. Therefore, in order to recognize and normalize genes and markers from texts, AlvisNLP and HunFLAIR both rely on curated domain lexicons or dictionaries combined with patterns. For wheat genes and markers, a curated list of gene names from the GrainGenes [29] database was created. For rice genes, the Oryzabase [7] database was used and integrated into the AgroLD Knowledge Graph [26] which capitalises genomic data about plant species of high interest for the plant science community (among which rice and wheat) to provide functional information on genes and their relationship across species. AgroLD is available through a SPARQL endpoint.[16]

For wheat varieties recognition and normalization, a curated list was created combining two sources: (1) the *Plant variety catalogues, databases & information systems*[17] and (2) the *Official Catalogue of Species and Varieties of Cultivated Crops*.[18] To be compliant with LOD principles, we created a URI to identify each distinct entry in the different created lexicons.

[13] The version 3.2 is available at https://agroportal.lirmm.fr/ontologies/CROPUSAGE.

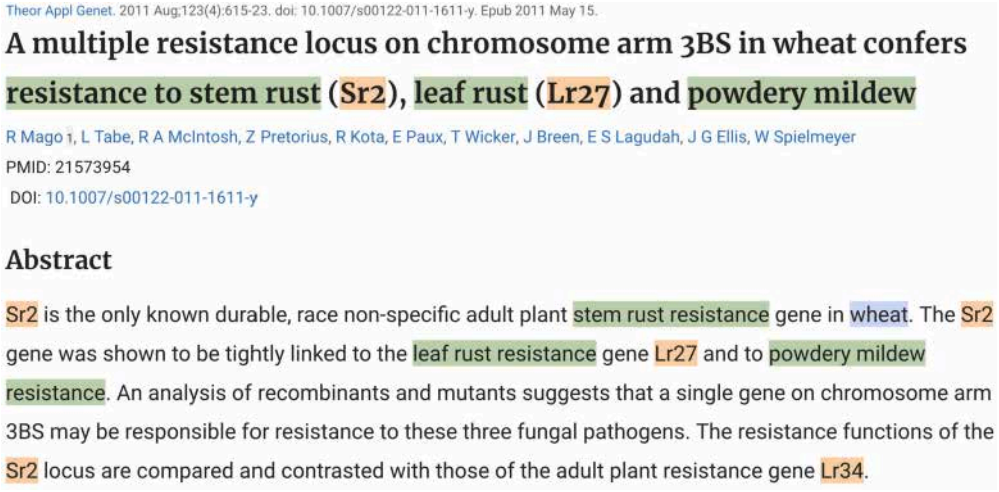
[14] The version 1.2 is available at https://agroportal.lirmm.fr/ontologies/PPDO.

[15] https://sparql.uniprot.org/sparql/.

[16] http://sparql.southgreen.fr.

[17] https://food.ec.europa.eu/plants/plant-reproductive-material/plant-variety-catalogues-databases-information-systems.

[18] https://www.geves.fr/catalogue/.


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 ī, 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.

Fig. 4. Example of NE recognition and linking in a PubMed publication.

3. Method

3.1. Competency questions

In this section, we present a set of Competency Questions (CQs) stemming from requirements expressed by experts and collected in the context of the D2KAB project.[19] CQs are natural language questions illustrating the typical knowledge that scientists would require a data source to provide. Each corpus had its own panel of associated experts interested in working on the corpus. We interviewed them to find out what information they wanted to extract from the corpus. A common way of validating a KG is to provide the formalisation of CQs, for a given case study, as SPARQL queries using the KG model. In the following we present the CQs for our case studies. Their formalisation in SPARQL is presented in Section 4.

3.1.1. Competency questions for scientific literature exploration in wheat and rice functional genomics

One of the most common investigated research questions in functional genomics are those related to genotype-phenotype relationships. However, they are not always straightforward to be identified. Considering rice and wheat genomes, they differ considerably in terms of size and complexity. The rice genome is relatively small compared to the wheat genome, comprising around 430 million base pairs in its haploid form. The wheat genome is much larger and more complex, with a hexaploid genome made up of three sets of chromosomes and comprising around 17 billion base pairs. Research in both rice and wheat functional genomics has already led to several important advances, such as the development of varieties with enhanced disease resistance and improved nutritional content.

Hence, exploiting the ever-growing scientific literature could help scientists to discover hidden interactions between entities of interest for functional genomics by examining their co-occurrence in scientific publications. Thus, structuring and integrating genomic NEs extracted from scientific publications and annotated based on relevant knowledge from external semantic resources is essential. Considering the PubMed publications corpus presented in Section 2, we present here a subset of CQ that ultimately consists of a set of research questions.

CQ1. Which genes are mentioned proximal to a specific trait (e.g., resistance to Fusarium head blight, resistance to leaf rust)?

CQ1 expresses the importance of supporting experts in identifying genetic entities recognized proximal to a particular trait in order to establish possible links between gene expressions and traits. For instance, CQ1 addresses the need of scientists to discover genes involved in the resistance to biotic or abiotic factors in both wheat and rice species based on scientific literature. As illustrated in Fig. 4, several gene names proximal to a given wheat trait are recognized in PubMed scientific publications. Thus, genes that are involved in resistance to a specific disease can be discovered on the basis of their presence next to specific disease-resistance traits within scientific literature. Taking the example of the *resistance to leaf rust*, there are several genes that have been identified as being associated with resistance to rust in wheat crops. The *Lr34* gene is a major gene for resistance to leaf rust. This type of knowledge is valuable in wheat breeding programs to develop varieties that are resistant to a specific disease.

CQ2: Which genetic markers appear proximal to a specific gene, and which genes are mentioned proximal to a particular phenotype in publications dating from after 2010?

The CQ2 is designed for the PubMed corpus on wheat genomics, since the NEs of the genetic markers are recognized only in this corpus. A genetic marker discriminates the different alleles of a gene with the polymorphism of the DNA sequence. Thus, genetic markers are used to select the wheat varieties with a trait or phenotype of agronomic interest [15]. For instance, in the case of *resistance to the stripe rust disease* in wheat, the gene *Yr65* is often mentioned in literature along with this phenotype. Furthermore, markers such as *Xgdm33*, *Xgwm11*, *Xgwm18*, and *Xgwm413* are mentioned in the same context as this gene. As the techniques for genetic markers selection have evolved over time and some of them have become obsolete, the expert can also refine the query to select only publications which appeared after 2010. The knowledge graph should contain the publication metadata such as publication year, list of authors, or the number of incoming citations.

CQ2-bis: Which chemical compounds are cited in scientific publications proximal to gene names, and which genes are in turn mentioned proximal to a particular phenotype?

Chemical compounds are often involved in metabolic processes which are controlled by genes. In scientific literature, associations between chemical compounds and genes can reveal interesting phenotypes. CQ2-bis emphasizes that biologists can search for rice genes that co-occur with a specific phenotype and a chemical compound.

CQ3. Which scientific publications mention gene names that appear proximal to a specific wheat or rice variety name and a trait from a specific given class of traits (e.g. all traits related to fungal pathogen resistance)?

CQ3 reflects the need for experts to conduct a systematic literature review of publications that mention specific genes cited in the literature proximal to certain traits (from a specific family of traits) as well as wheat or rice varieties. The results of this query should include a list of articles mentioning, in their abstracts or titles, gene names, a wheat or rice variety and a set of traits known, for instance, to be involved in

[19] https://www.d2kab.org/.

pathogen resistance. For instance, a scientist may be interested in resistance to fungal pathogens which cause massive and destructive losses to crops. Thus, the study of resistance mechanisms is essential to fully understand the interactions between pathogens across crop varieties. Based on the WTO structure which classifies traits in different taxonomies, it is possible to conduct this study for all traits belonging to the sub-hierarchy of fungal pathogen resistance class. This CQ highlights the importance to incorporate domain knowledge formally represented in ontological and terminological resources (e.g., WTO).

CQ4: Which gene names are cited in the literature proximal to a specific taxon (and optionally to one or more of its descendants)?

CQ4 reflects the need to perform a search of gene mentions cited proximal to different taxa mentions. We may initiate the query by focusing on a single taxon mention and expand it dynamically by including each descendant taxon. So, the query shows first results for a single search on a specific taxon mention. Then, it generates a more comprehensive set of results.

CQ5: What are orthologous genes in rice and wheat genomes?

It has been demonstrated that some fungal and bacterial disease pathogens affect both rice and wheat. Wheat and rice disease resistance has been studied for a large panel of pathogens, including *rusts, smuts, Fusarium head blight, Septoria leaf blotch, tan spot, and powdery mildew*, that cause the most serious losses. The goal is to search for wheat and rice genes co-occurring in literature with the same taxon of a pathogen (or a more specific taxon). This enables to identify orthologous genes[20] in wheat and rice.

3.1.2. Competency questions for agronomic studies

Climatic change has an impact on agriculture practices. Agronomists would like to study PHBs in order to analyse the distribution of crops on the French territory and provide answers to several questions such as: have the farmers changed the crops they produce over the time? In addition, agronomists would like to study how climatic change has affected crop growth. Indeed, due to variable weather conditions, crop development can differ from year to year. One of the mid-term objectives of the D2KAB research is to create a timeline of the development stages of crops in specific regions of France. As each PHB is related to a unique region of France and has a publication date, by extracting the crop names from PHB text, it is possible to identify the crops to which the PHB relates, and thus determine which crops were grown in that region at a given time. Extracting crop development stages from the PHB text also allows experts to understand the development stage that the crop had reached at the time of publication in that region. Considering the PHB corpus presented in Section 2, we present here a subset of CQ that ultimately consists of a set of research questions.

CQ6: Which crop names are mentioned in the title of a specific PHB?

This CQ aims to identify the topic of the PHB, i.e., the main crop or crop category mentioned in one of the titles of the PHB. A PHB title may mention one or several crop names. In the example of Fig. 2, the term *Viticulture* is mentioned in the title, thus the PHB is about a single crop which is cultivated grapevine.

CQ6 bis: How many times is a crop name mentioned in a specific PHB?

The goal is also to identify the main crops or crop categories that represent the topic of a PHB, thus reinforcing the previous CQ. One way to identify the main crop topic of a PHB is to count how many times a crop name appears in the text of a PHB. The crop mention may appear in any type of section (e.g. footer).

CQ7: What are the most cultivated crops in a given French region and do they change over time?

The scientific objective is to find which crops are cultivated in a specific region of France. Based on the PHB corpus, it is possible to retrieve the subset of bulletins concerning a French region for a specific period of time. Then, we can compute the main crop topics of these bulletins.

CQ8: Which development stages are mentioned proximal to a crop name in a specific PHB?

The goal is to identify when a development stage of a specific crop is observed in a specific region, given that the crop name, development stage may be in separate paragraphs of a PHB.

In the example of Fig. 2, the crop name is mentioned in the title and several development stages are mentioned in the first paragraph of the middle column.

CQ9: What is the scientific literature available on the crop to which a PHB bulletin relates? The goal is to identify which new research publications are related to a crop cultivated in the French territory, to search for example new crop varieties that are resistant to drought or high temperatures.

Table 1
List of reused vocabularies.

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/
prov	http://www.w3.org/ns/prov#
frbr	http://purl.org/vocab/frbr/core#
obo	http://purl.obolibrary.org/obo/
d2kab_inrae	http://ontology.inrae.fr/bsv/ontology/
d2kab	http://ns.inria.fr/d2kab/
dul	http://www.ontologydesignpatterns.org/ont/dul/DUL.owl#
frac	http://www.w3.org/ns/lemon/frac#

3.2. Proposed semantic model

We reuse a set of state-of-the-art vocabularies to design a unified semantic model that captures the context of occurrence of several types of NE annotations in documents. The core part of this model leverages and extends the model previously proposed in [12]. It is based on the W3C Web Annotation Ontology (OA) [23] to structure, describe and integrate NEs extracted from both corpora, and eight complementary vocabularies to describe documents and NEs. Table 1 shows the main vocabularies used to describe named entities annotations as well as documents in both corpora.

3.2.1. OA-based model for text annotations with named entities

The Web Annotation Data Model is an ontology for structuring and sharing any type of annotations in an interoperable format. According to the OA documentation,[21] "*an annotation is considered to be a set of connected resources (each identified by a URI), typically comprising a body and a target where the body is somehow about the target*". The core OA data model is that an annotation a_i is an instance of the `oa:Annotation` class such that:

- The `oa:hasTarget` property identifies the part of document that is being annotated with annotation a_i. The target is a resource selection with a selector, i.e., a resource that identifies the part of text m_e that mentions a recognized entity e. In this work, we use different types of selectors: `oa:TextQuoteSelector`, `oa:TextPositionSelector`, `oa:XPathSelector` to indicate respectively the NE's mention m_e (i.e., surface form), the start and end offset position of m_e in the text and/or the XPath expression to retrieve m_e in the HTML element. Note that only PHB corpus contains HTML documents. Therefore, oa:XpathSelector is only used for annotations

[20] found in different organisms, but derived from a single common ancestral gene present in the common ancestor of those organisms.

[21] https://www.w3.org/TR/annotation-model/.

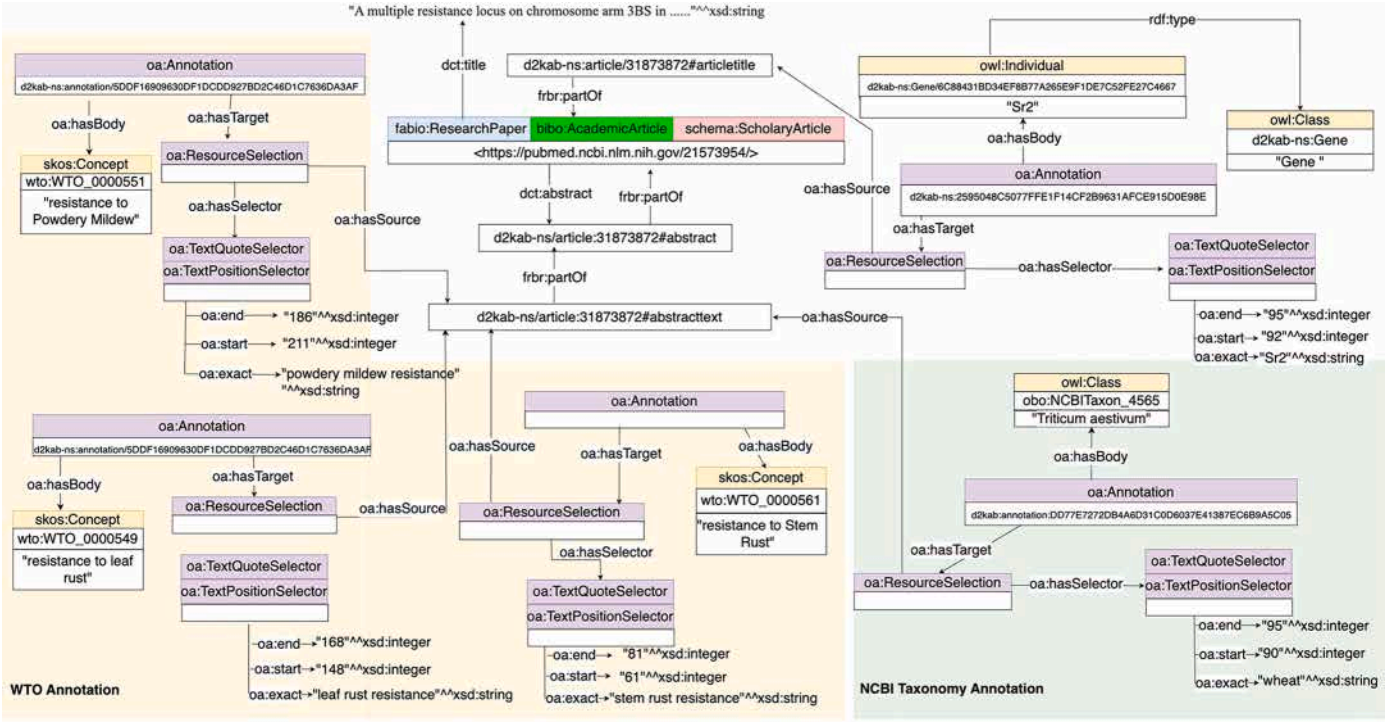


Fig. 5. Example of NE annotations identified in a PubMed Publication's section (title and abstract) and represented in WheatGenomicsSLKG based on OA ontology.

extracted from this corpus. The `oa:hasSource` property is used to specify the URI of the source where the selector is applied, the source being either the URI of the document or one of its sub-parts.
- The `oa:hasBody` property identifies the entity e defined in a domain vocabulary such as WTO, NCBI taxonomy, PPDO or FCU thesaurus.

Fig. 5 illustrates an example RDF graph that captures five instances of NE annotations recognised in the title and the abstract of a publication in the PubMed corpus.[22] The title and the abstract of the publication are identified by a URI and become the source of the target selector. Three annotations have as body a SKOS concept in the WTO resource (yellow area in Fig. 5). One annotation has as body a class from the NCBI taxonomy. One annotation has as body a gene entity URI that we have created locally in our graph (green area in Fig. 5). All mentions are identified by two selectors:

- an instance of `oa:TextQuoteSelector` is used to specify the text of the mention.
- an instance of `oa:TextPositionSelector` is used to specify the start and end offset position of the mention.

Fig. 6 represents three annotations extracted from the PHB[23] presented in Fig. 2.

One annotation[24] identifies the mention *Viticulture* localized in the main title of the PHB. Three types of selectors are used:

- an instance of `oa:XpathSelector` is used to express that the mention is found in the first section of the HTML element of type H1 which is a first level title.
- an instance of `oa:TextQuoteSelector` is used to specify the text of the mention, its prefix and suffix.
- an instance of `oa:TextPositionSelector` is used to specify the start and end offset position of the mention.

The annotation body is a SKOS concept from FCU thesaurus.

Note that a mention found in a text may concern several entities. For instance the second annotation in Fig. 6 identifies in the mention *BBCH-69 et 73* two entities from PPDO: the development stages *BBCH 69* and *BBCH 73*. Thus two distinct annotations, associated to two distinct bodies share the same resource selection.

The `oa:motivatedBy` property identifies the motivation of the annotation creation. Since all annotations a_i aim to identify an entity e in the text of the document, the object of this property is `oa:identifying` which is an instance of class `oa:Motivation`.

3.2.2. Bibliographic metadata

To describe bibliographic metadata of documents in the corpora, we have reused the following vocabularies: Dublin Core,[25] FRBR aligned bibliographic ontology (FaBiO) [19], bibliographic ontology (BIBO),[26] Dolce Ultra Light (DUL) [18], PROV Ontology (PROV) [11], the module for FRequency, Attestation and Corpus information (FRAC) of the LExicon Model for ONtologies (LEMON) [3] and Schema.org. These vocabularies have been used slightly differently for each corpus.

Bibliographic metadata of PubMed scientific documents For the PubMed corpus, we have reused bibliographic metadata vocabularies to describe specific attributes of scientific documents such as DOI, year of publication, number of pages, journal, etc. First, a scientific article is represented as an instance of classes `fabio:ResearchPaper`, `bibo:AcademicArticle` and `schema:ScholarlyArticle`. The Dublin Core properties `dct:title` and `dct:abstract` link the document to its title and abstract. Note that, in the PubMed corpus, abstracts may be struc-

[22] https://pubmed.ncbi.nlm.nih.gov/21573954/.
[23] http://ontology.inrae.fr/bsv/resources/Q16994/2019/bsv_viti_13_27_06_2019_cle07f426_html.
[24] http://ontology.inrae.fr/bsv/resources/Q16994/2019/bsv_viti_13_27_06_2019_cle07f426/aa_230327/VITICULTURE_FCU_1_sel.
[25] https://www.dublincore.org/specifications/dublin-core/dcmi-terms/.
[26] https://github.com/structureddynamics/Bibliographic-Ontology-BIBO.

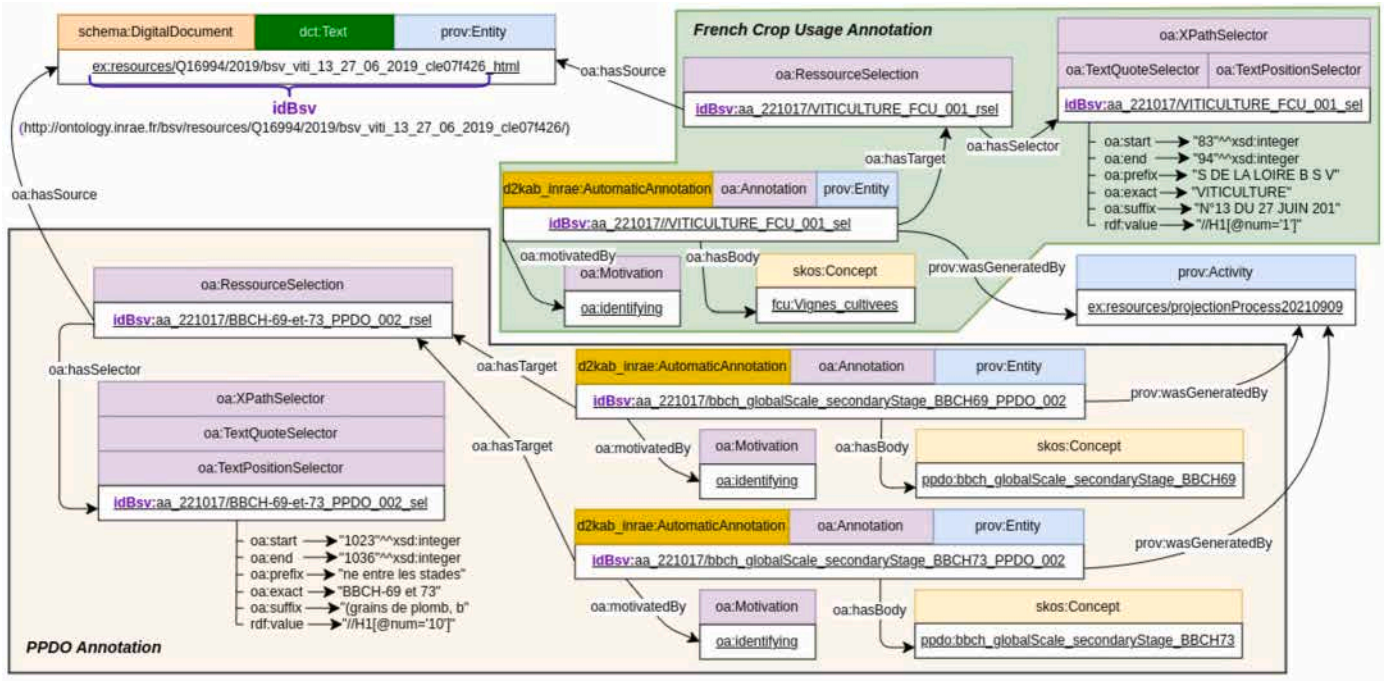


Fig. 6. Example of NE annotations identified in a PHB. One annotation identifies crop, and the other one identifies two development stages.

tured in three subsections distinguished in our model by three different resources, each one identified by a unique URI. Property `frbr:partOf` is used to link an abstract and the document it is related to, or an abstract and one of its sub-sections. Fig. 7 illustrates (a subset of) the bibliographic metadata of a scientific document.

Bibliographic metadata for PHB technical documents In the PHB corpus, a bulletin has two digital realizations: a PDF file and a HTML file. Therefore, to model the bibliographic information, we have reused an ontology design pattern from the DUL ontology called `dul:InformationObject` [18]. An information object represents the generic information about a document such as its publication date, the corpus to which it belongs, its associated French region, its description. An information object has several realizations represented by using property `dul:isRealizedBy`.

Fig. 8 presents the graph annotating the bulletin of Fig. 2.

The bulletin is an instance of class `d2kab_inrae:Bulletin` which specializes `dul:InformationObject`. The Dublin Core properties `dct:date`, `dct:description` and `dct:spatial` link the bulletin to its publication date, its description accessible on the download page, its French region extracted from the download web site and identified by its wikidata URI. Each sub-corpus is represented by an instance of `prov:Collection`. A bulletin belongs to at least one sub-corpus which is represented by using property `prov:hasMember`. The files are instances of classes `schema:DigitalDocument` and `dct:Text`. The Dublin Core properties `dce:language` and `dce:format` link a file to its language and format. The OA property `oa:textDirection` links a file to its text direction. The property `schema:url` links a file to its URL where it is actually accessible. The property `schema:isBasedOn` links a file to the URL where it was previously downloaded. The property `frac:total` links the HTML file to its total number of tokens.

3.2.3. Provenance metadata

The FCU thesaurus has evolved over time and several versions of it exist. Moreover different NLP processes based on different versions of FCU were tested on the PHB corpus to generate annotations. The PROV ontology is used to store the provenance information of the annotations. An example provenance metadata of a PHB annotation is shown on Fig. 9. Each instance of `oa:Annotation` is linked to an instance of `prov:Activity` which generated it. Properties `prov:startedAtTime` and `prov:endedAtTime` link the activity to the date when the NLP pipeline was applied on the sub-corpus. Property `prov:used` indicates the version of FCU thesaurus. Regarding the activity, property `prov:qualifiedAssociation` indicates the NLP pipeline plan and the NLP software used to run the plan. Regarding the plan, properties `prov:wasAttributeTo`, `prov:generatedAtTime`, and `schema:url` indicate its author, its creation date and its git repository.

3.3. Data transformation pipeline

To create the three KG, we adopted a materialization approach in which mapping rules are defined to transform raw annotations generated by NLP pipelines into RDF. We relied on the xR2RML mapping language [8] to define the mapping rules that formally describe the relationship between raw annotations, initially stored in CSV files, and classes and properties from the semantic model. The translation was carried out by an implementation of xR2RML for MongoDB databases, Morph-xR2RML.[27] Each mapping rule defines a Triple Map (`rr:TripleMap`) which expresses a generic pattern for generating RDF triples in accordance with the model proposed in Section 3.2. Fig. 10 summarizes all the steps of our methodology. It makes clear that different tools can be used for NLP (pre)-processing. The mapping-based transformation step is common to all graphs construction pipelines.

3.3.1. KG pipeline for scientific literature on wheat and rice genomics

The xR2RML mapping rules defined to materialize the knowledge graph describing the scientific literature on wheat genomics, WheatGenomicsSLKG, are available in the project's GitHub directory.[28] Similar

[27] https://github.com/frmichel/morph-xr2rml/.

[28] https://github.com/Wimmics/d2kab-wheatGenomicsLiterature-kg/tree/main/mapping-rules.

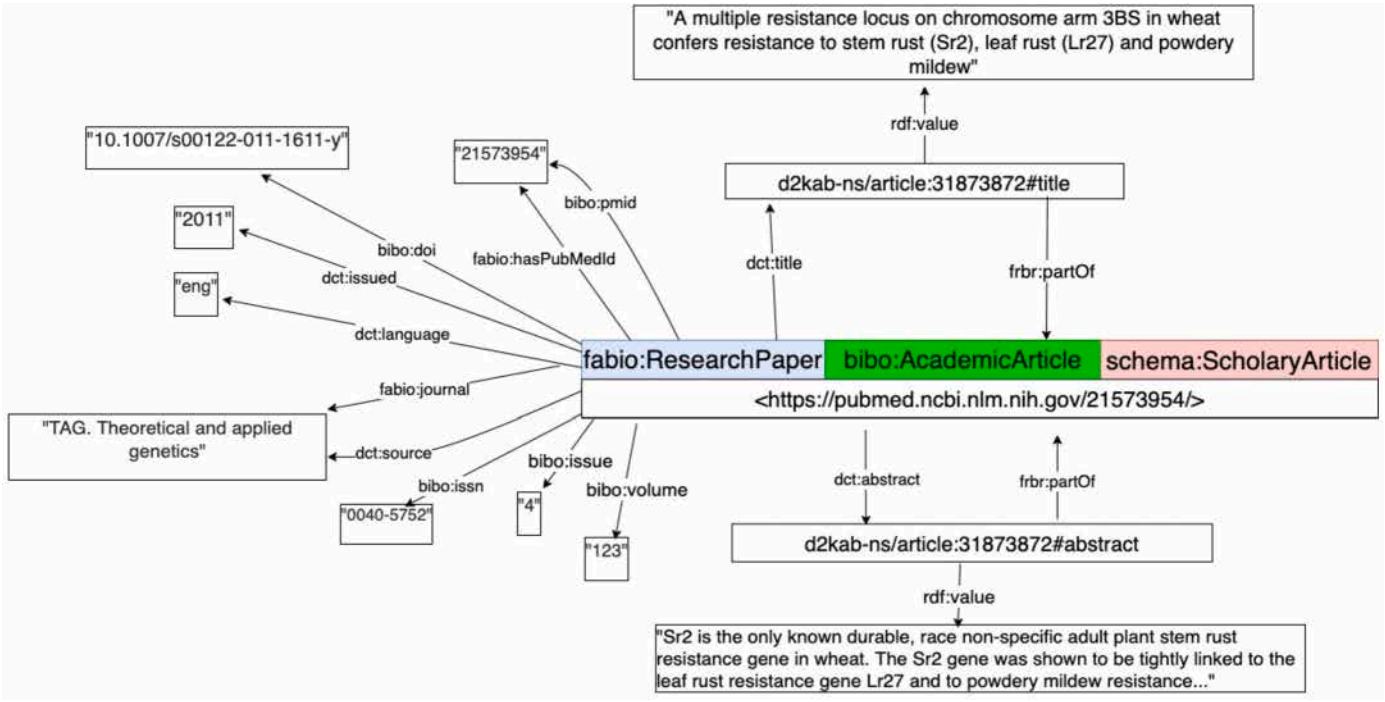


Fig. 7. An RDF graph describing bibliographic metadata of a scientific publication in the PubMed corpus.

Table 2
Templates of URI for the resources in WheatGenomicsSLKG and RiceGenomicsSLKG.

	URI Template
Entity	http://ns.inria.fr/d2kab/{EntityClass}/{EntityID}
Article	https://pubmed.ncbi.nlm.nih.gov/{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}

mapping rules have been defined to materialize the knowledge graph describing the scientific literature on rice genomics, RiceGenomicsSLKG; they are available in the project's GitHub directory.[29] Table 2 illustrates the templates used to generate significant URIs for different types of resources in WheatGenomicsSLKG and RiceGenomicsSLKG.

In addition, in order to enrich the scientific publications with bibliographic metadata, we developed a SPARQL micro-service [10] to query the PubMed Central API and retrieve publication metadata.[30] For each publication, the micro-service transforms PubMed API's results into an RDF graph that we insert in the KG being constructed. Finally, we also inserted as a subgraph of WheatGenomicsSLKG the SKOS version of the WTO semantic resources used to annotate phenotypes entities.

3.3.2. KG pipeline for plant health bulletins

Since the beginning of their publications, PHBs have been made freely available in PDF format on the websites of the Regional Chambers of Agriculture or the websites of the regional agency of the French Ministry of Food and Agriculture (DRAAF). Therefore, PHBs are disseminated on different websites (one per region).

A web-crawler is periodically run over the DRAAF websites to look for new PHBs that are downloaded while some information is extracted (download date, download URL, local filename and web path) [21]. These data are transformed into RDF using python scripts. The downloaded pdf files are transformed into HTML using the `pdf2blocks`[31] conversion tool. AlvisNLP pipelines are used to extract NEs from these HTML files. Finally, the CSV output files are transformed using specific xR2RML mapping rules. All the elements of this workflow are available in the project gitlab repository.[32]

4. Results and validation

In this section, we first present the obtained graphs, followed by the implementation of the various CQs presented in section 3.1. All queries are available on our git repositories and different notebooks are provided showing examples of results. All details are provided in section 4.2.

4.1. Resulting knowledge graphs

We have built three different knowledge graphs considering three distinct agricultural corpora. Tables 3 and 4 describe key statistics of built KGs.

4.2. Implementation of competency questions

In order to demonstrate how the three KG serve several expert needs, we implemented the competency questions presented in Section 3.1 in SPARQL. All the presented CQ could be translated into SPARQL queries and their results analysed as valid, which shows that our semantic model fulfills the requirements. It is worth noticing that, although different classes of NE are recognized in the different corpora, the structure of SPARQL queries is quite similar.

[29] https://github.com/ANR-DIG-AI/RiceGenomicsSLKG.
[30] https://sparql-micro-services.org/service/pubmed/getArticleByPMId_sd/.
[31] https://doi.org/10.5281/zenodo.4067965.
[32] https://forgemia.inra.fr/stephan.bernard/corpus-bsv.

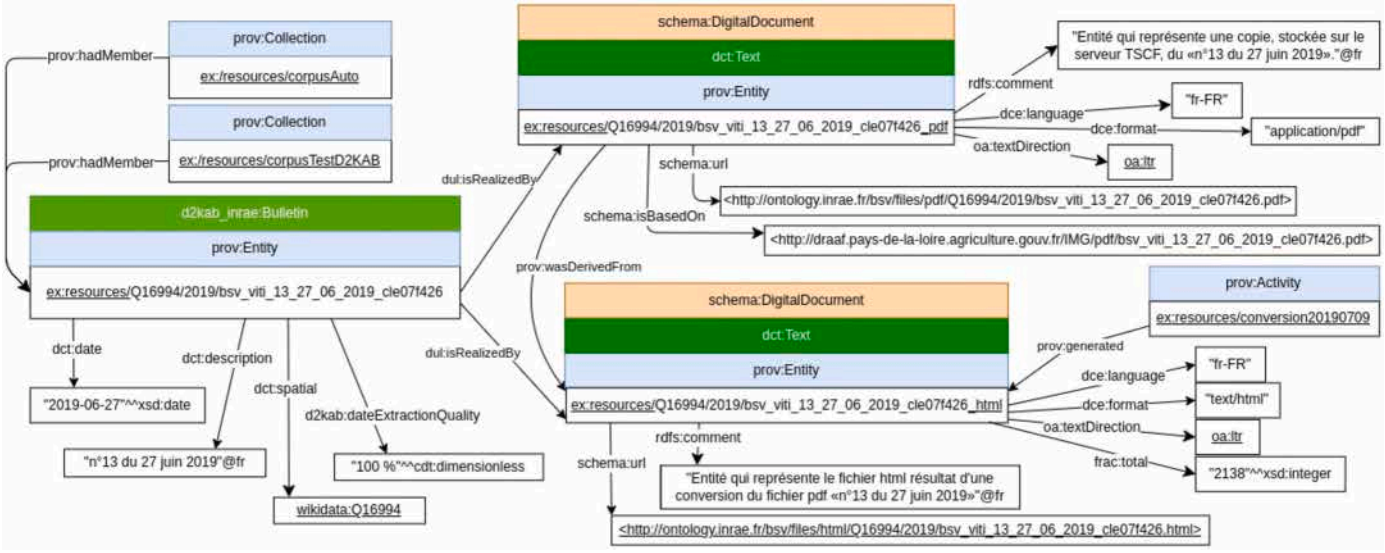


Fig. 8. Example RDF graph describing metadata of a bulletin in the PHB corpus.

Table 3
Key statistics of the WheatGenomicsSLKG and RiceGenomicsSLKG Knowledge Graphs.

SPARQL endpoint → http://d2kab.i3s.unice.fr/sparql			
WheatGenomicsSLKG		RiceGenomicsSLKG	
Number of RDF Triples	1.191.867	Number of RDF Triples	3.971.995
Total number of annotations	88.879	Total number of annotations	348.089
Total number of Pubmed abstracts	8.496	Total number of Pubmed abstracts	17.627
Distribution of annotations per NE class			
Gene	11.292	Gene	40.789
Genetic Marker	932	Chemical	9.546
Wheat Trait	4.630	Phenotype/diseases	1.808
Taxon	71.833	Taxon	5.469
Total number of varieties	192		

Table 4
Key statistics of the PHB Knowledge Graph.

SPARQL endpoint → http://ontology.inrae.fr/bsv/sparql	
Number of RDF Triples	1.352.654
Total number of annotations	75.522
Total number of PHB	1.074
Distribution of annotations per NE Class	
Crop identification	59.127
Growth stage	16.041
Distribution of annotations per sub-corpora	
Vespa Corpus	
28.449 crop annotations	6.516 growth stage annotations.
497 PHB	355 NEs (296 crop + 59 growth st.)
D2KAB Corpus	
16.728 crop annotations	5.708 growth stage annotations.
230 PHB	330 NEs (224 crop + 106 growth st.)
Alea Corpus	
12.565 crop annotations	3.027 growth stage annotations.
150 PHB	326 NEs (237 crop + 89 growth st.)

4.2.1. *SPARQL queries implementing CQs on the wheat and rice PubMed corpora*

The SPARQL queries implementing CQs on PubMed corpora of wheat and rice functional genomics can be executed at the SPARQL endpoint.[33] The queries and excerpt of the obtained results are provided as part of the supplementary materials. A Jupyter Notebook of these SPARQL queries is available on our github repository.[34]

CQ1: The SPARQL query presented in Listing 1 implements CQ1 and allows scientists to retrieve genes that are mentioned proximal to the *resistance to leaf rust* trait considering the WheatGenomicsSLKG graph. The query returns all genes mentioned proximal to the WTO concept (`wto:0000483`) that corresponds to the aforementioned trait and counts the number of times that a gene and the trait are recognized in the same context. The results of this query confirm that Lr34 is the most cited gene in the literature. Lr10, Lr26 and Lr24 genes appear also as the most frequent genes.

CQ2 and **CQ2-bis**: The SPARQL query presented in Listing 2 implements CQ2 and allows to identify genetic markers and genes mentioned proximal to a specific wheat trait in scientific publications. The results of this query return a list of scientific publications from the WheatGenomicsSLKG graph that list several genetic markers and genes entities

[33] http://d2kab.i3s.unice.fr/sparql.
[34] https://github.com/Wimmics/d2kab-wheatGenomicsLiterature-kg/blob/main/SPARQLQueries-JupyterNotebook.ipynb.

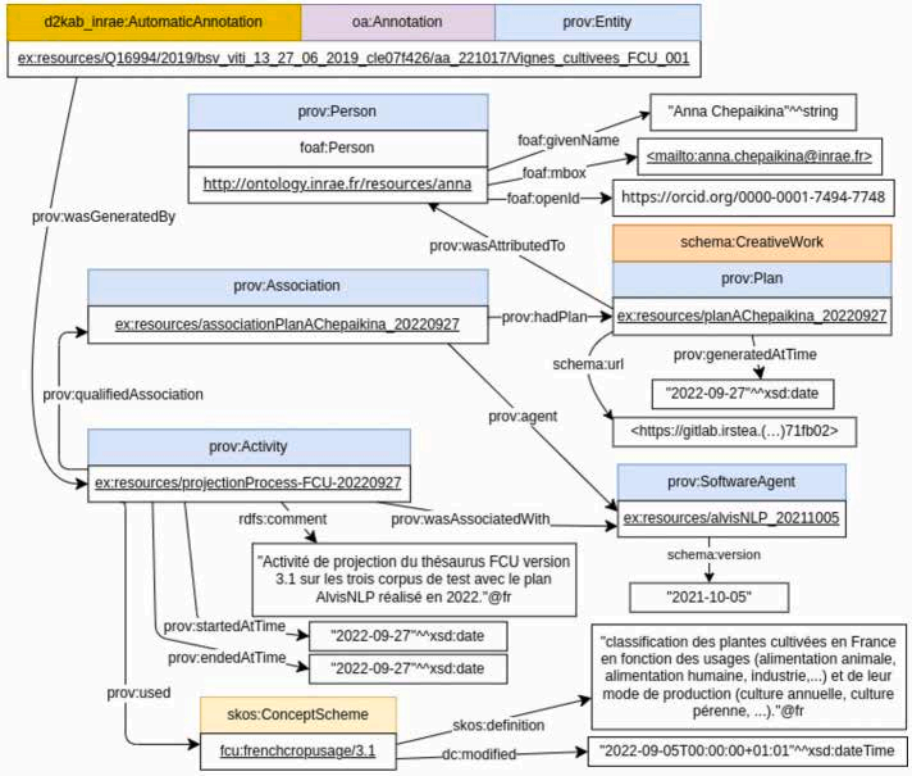


Fig. 9. Example RDF graph describing provenance metadata of a bulletin in the PHB corpus.

```
1  SELECT ?GeneName (count(distinct ?paper) as ?NbOcc)
2  FROM NAMED <http://ns.inria.fr/d2kab/graph/wheatgenomicsslkg>
3  FROM NAMED <http://ns.inria.fr/d2kab/ontology/wto/v3>
4  WHERE {
5    GRAPH <http://ns.inria.fr/d2kab/graph/wheatgenomicsslkg> {
6      ?a1 a oa:Annotation;
7          oa:hasTarget [ oa:hasSource ?source1 ] ;
8          oa:hasBody ?WTOtraitURI .
9      ?source1 frbr:partOf+ ?paper .
10     ?a a oa:Annotation ;
11       oa:hasTarget [ oa:hasSource ?source ] ;
12       oa:hasBody [ a d2kab:Gene; skos:prefLabel ?GeneName ] .
13     ?source frbr:partOf+ ?paper .
14     ?paper a fabio:ResearchPaper . }
15   GRAPH <http://ns.inria.fr/d2kab/ontology/wto/v3> {
16     ?WTOtraitURI skos:prefLabel "resistance to Leaf rust" . }
17 }
18 GROUP BY ?GeneName
19 HAVING (count(distinct ?paper) > 1)
20 ORDER BY DESC(?NbOcc)
```

Listing 1: SPARQL query implementing CQ1 and retrieving the most cited genes mentioned proximal to the "*resistance to Leaf rust*" trait in WheatGenomicsSLKG (execution time <0.2 s).

mentioned proximal to the *resistance to Stripe Rust* trait. On another side, the SPARQL query presented in Listing 3 corresponds to the implementation of CQ2-bis and allows scientists to retrieve gene names that are mentioned proximal to the *GDP* chemical component in the scientific literature on rice genomics.

CQ3: The SPARQL query, presented in Listing 4, implements CQ3 and allows scientists to retrieve publications in which genes are mentioned proximal to wheat varieties and traits from a specific class, e.g., all wheat traits related to resistance to fungal pathogens. Based on the WTO structure which classifies traits in different taxonomies, the query retrieves all traits belonging to the sub-hierarchy of fungal pathogen resistance class (line 20-45).

CQ4: A first implementation of this CQ is presented in Listing 5 that performs a search of gene mentions cited proximal to a specific taxon identified by a class in the NCBITaxon ontology. Different taxa mentions can be also identified proximal to genes mentions in scientific publications in both wheat and rice corpora. The SPARQL query presented in Listing 6 extends the search for all sub-classes of a specific NCBITaxon class (*Puccina*[35]).

4.2.2. SPARQL queries implementing CQs on the PHB corpus

The SPARQL queries implementing CQ on the PHB corpus can be executed at the endpoint http://ontology.inrae.fr/bsv/sparql. Since not all FCU crop concepts have English labels, the results of the queries may vary depending on the label language. In the following SPARQL queries, only French labels are requested. A Jupyter Notebook of these SPARQL queries is available on our Github repository.[36]

CQ6: The SPARQL query presented in Listing 7 implements CQ6 and retrieves all the crop names that are mentioned in the H1 section of the HTML versions of PHBs. Only 618 bulletins out of 880 have a crop annotation in their H1 sections and some bulletins have several crop names identified within them (max 24).

CQ6 bis: A variation of the SPARQL query implementing CQ6 is presented in Listing 8 that retrieves the number of times that a crop name appears in a specific PHB (Fig. 2 PHB example). The result shows that the most recognized FCU concept is grapevine (appearing six times). Eight distinct FCU concepts are recognized in the text of this PHB.

CQ7: The SPARQL query presented in Listing 9 implements CQ7 and retrieves the number of times that each crop name is mentioned in the

[35] http://purl.obolibrary.org/obo/NCBITaxon_5296.
[36] https://forgemia.inra.fr/bsv/corpus-bsv/-/blob/SAAD/sample/PHB-KG_SPARQL_Queries.ipynb.

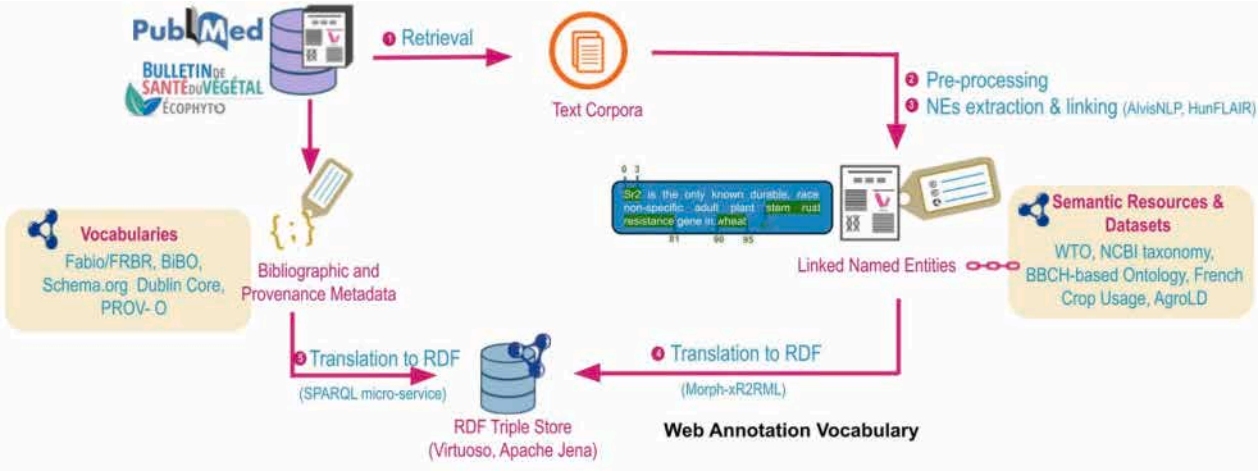


Fig. 10. The overall pipeline of KGs construction methodology.

```
SELECT (GROUP_CONCAT(distinct ?GeneName; SEPARATOR="-") as ?genes)
  (GROUP_CONCAT(distinct ?marker; SEPARATOR="-") as ?markers) ?paper ?year ?WTOtrait
FROM NAMED <http://ns.inria.fr/d2kab/graph/wheatgenomicsslkg>
FROM NAMED <http://ns.inria.fr/d2kab/ontology/wto/v3>
WHERE {
  VALUES ?WTOtrait { "resistance to Stripe rust" }
  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 [ a d2kab:Marker ; skos:prefLabel ?marker ] .
    ?source2 frbr:partOf+ ?paper .
    ?a3 a oa:Annotation ;
        oa:hasTarget [ oa:hasSource ?source3 ] ;
        oa:hasBody ?WTOtraitURI .
    ?source3 frbr:partOf+ ?paper .
    ?paper a fabio:ResearchPaper ; dct:title ?source3 ; dct:issued ?year .
    FILTER (?year >= "2010"^^xsd:gYear) }
  GRAPH <http://ns.inria.fr/d2kab/ontology/wto/v3> {
    ?WTOtraitURI skos:prefLabel ?WTOtrait . }
}
GROUP BY ?paper ?year ?WTOtrait
```

Listing 2: SPARQL query implementing CQ2 and retrieving genes names and genetic markers mentioned proximal to the wheat trait *resistance to Stripe Rust* (execution time ≈ 1.42 s).

```
SELECT ?GeneName (count(distinct ?paper) as ?NbOcc)
WHERE {
  ?a1 a oa:Annotation;
     oa:hasTarget [ oa:hasSource ?source1 ] ;
      oa:hasBody [ a d2kab:Chemical; skos:prefLabel "GDP"] .
  ?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 .
}
GROUP BY ?GeneName
HAVING (count(distinct ?paper) > 0)
ORDER BY DESC(?NbOcc)
```

Listing 3: SPARQL query implementing CQ2-bis and retrieving gene names that are mentioned proximal to the *GDP* chemical component in RiceGenomicsSLKG (execution time < 0.20 s).

subset of bulletins for the French region *Pays de la Loire*, ordered by descending order. It estimates the most cultivated crops in this region during the whole time period. This query could also include a specific time period to observe the evolution of cultivated crop in the region. The results show that grape, cabbage, leek and carrot are the most cultivated crops in *Pays de la Loire*. These are more precise results than those from our previous work [20] in 2017 indicating that this region growth field crops, vegetables and fruits.

CQ8: The SPARQL query presented in Listing 10 implements CQ8 and retrieves couples of annotations, one for a crop name and one for a development stage, that are localized in the same HTML element of a PHB. This query then estimates that the development stage is applicable to the crop. Thus one can deduce that at the publication date of the PHB the crop has reached the development stage in the region the PHB is relative to. The query retrieves 1304 HTML elements that contain both an annotation of FCU crop concepts and of PPDO development stages from 190 distinct PHBs. Note that the execution of this query takes some time (50 s) due to the call of two service templates.

```
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/0000340> . }
    UNION
    { ?body rdfs:label ?WTOtrait ;
        rdfs:subClassOf* <http://opendata.inrae.fr/wto/0000340> . }
    UNION
    { ?body skos:prefLabel ?WTOtrait ; skos:broader* ?concept .
      ?concept a ?class .
      ?class rdfs:subClassOf* <http://opendata.inrae.fr/wto/0000340> . } }
} LIMIT 20
```

Listing 4: SPARQL query implementing CQ3 and retrieving all genes cited proximal to wheat varieties and traits from a specific family of traits (execution time ≈ 1.55 s).

```
SELECT distinct ?paper ?title (GROUP_CONCAT(distinct ?geneName; SEPARATOR="-") as ?genes)
  ?ncbiTaxon
FROM NAMED <http://ns.inria.fr/d2kab/graph/wheatgenomicsslkg>
FROM NAMED <http://purl.obolibrary.org/obo/ncbitaxon/ncbitaxon.owl>
WHERE {
  VALUES ?ncbiTaxonURI {<http://purl.obolibrary.org/obo/NCBITaxon_208348>}
  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 .
    ?a3 a oa:Annotation ;
      oa:hasTarget [ oa:hasSource ?source2 ] ;
      oa:hasBody ?ncbiTaxonURI .
    ?source2 frbr:partOf+ ?paper .
    ?paper a fabio:ResearchPaper; dct:title ?titleURI .
    ?titleURI rdf:value ?title . }
  GRAPH <http://purl.obolibrary.org/obo/ncbitaxon/ncbitaxon.owl> {
    ?ncbiTaxonURI rdfs:label ?ncbiTaxon . }
} LIMIT 100
```

Listing 5: SPARQL query implementing CQ4 and performing a search of gene mentions proximal to a specific taxon in the NCBI Taxon ontology (execution time < 0.53 s).

In the PHB example of Fig. 2, the crop name and the development stage are mentioned in distinct HTML elements. The crop name is mentioned before the development stage. Another implementation consists in retrieving the crop name and the development stage in a window of a fixed number of characters. The following query retrieves the couples of annotations, one for the crop name and one for the development stage, that appear in the characters window of 1000 characters in the PHB example. This alternative implementation of CQ8 is presented in Listing 11.

4.2.3. *Combined exploitation of knowledge graphs*

Using federated queries, scientists can jointly exploit several KGs. In the following, we present an example combined exploitation of WheatGenomicsSLKG and RiceGenomicsSLKG and an example combined exploitation of WheatGenomicsSLKG and PHB KG.[37]

[37] Both queries are also available in the Jupiter notebook https://github.com/Wimmics/d2kab-wheatGenomicsLiterature-kg/blob/main/SPARQLQueries-JupyterNotebook.ipynb.

```
SELECT distinct ?paper ?title (GROUP_CONCAT(distinct ?geneName; SEPARATOR="-") as ?genes) ?ncbiTaxon
FROM NAMED <http://ns.inria.fr/d2kab/graph/wheatgenomicsslkg>
FROM NAMED <http://purl.obolibrary.org/obo/ncbitaxon/ncbitaxon.owl>
WHERE {
  VALUES ?ncbitaxonURI {<http://purl.obolibrary.org/obo/NCBITaxon_5296>}
  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 .
    ?a3 a oa:Annotation ;
      oa:hasTarget [ oa:hasSource ?source2 ] ;
      oa:hasBody ?ncbiTaxonURI .
    ?source2 frbr:partOf+ ?paper .
    ?paper a fabio:ResearchPaper ; dct:title ?titleURI .
    ?titleURI rdf:value ?title . }
  GRAPH <http://purl.obolibrary.org/obo/ncbitaxon/ncbitaxon.owl> {
    ?ncbiTaxonURI rdfs:subClassOf* ?ncbitaxonURI ; rdfs:label ?ncbiTaxon . }
} LIMIT 20
```

Listing 6: SPARQL query providing another implementation of CQ4 involving the hierarchical structure of the NCBI Taxon ontology - loaded as part of our knowledge graph (execution time ≈ 1.38 s).

```
SELECT ?phbDescription ?cropName ?xpath ?phbUri WHERE {
  SERVICE <http://ontology.inrae.fr/frenchcropusage/sparql> { GRAPH fcu:3.1 {
      ?body a skos:Concept ; skos:prefLabel ?cropName .
      FILTER (LANG(?cropName)='fr'). }}
  ?phbUri a d2kab_inrae:Bulletin ; dul:isRealizedBy ?phbHtml ; dct:description ?phbDescription.
  ?phbHtml dce:format "text/html" .
  ?rs a oa:ResourceSelection ; oa:hasSelector ?sel ; oa:hasSource ?phbHtml .
  ?aa a oa:Annotation ; oa:hasTarget ?rs ; oa:hasBody ?body .
 ?sel a oa:XPathSelector ; rdf:value ?xpath .
 FILTER contains(?xpath, "/H1[").
} LIMIT 50
```

Listing 7: SPARQL query implementing CQ6 and retrieving the crop names appearing in H1 sections of the HTML versions of PHBs *(execution time* ≈ 10*s)*.

```
SELECT ?cropName (count(?cropName) AS ?nb) WHERE {
<http://ontology.inrae.fr/bsv/resources/Q16994/2019/bsv_viti_13_27_06_2019_cle07f426>
      a d2kab_inrae:Bulletin ; dul:isRealizedBy ?phbHtml .
   ?phbHtml dce:format "text/html" .
   ?rs a oa:ResourceSelection ; oa:hasSource ?phbHtml .
   ?aa a oa:Annotation ; oa:hasTarget ?rs ; oa:hasBody ?body .
   SERVICE <http://ontology.inrae.fr/frenchcropusage/sparql> { GRAPH fcu:3.1 {
      ?body a skos:Concept ; skos:prefLabel ?cropName .
      FILTER (LANG(?cropName)='fr') }}
} GROUP BY ?cropName ORDER BY DESC(?nb)
```

Listing 8: SPARQL query implementing CQ6 and computing the number of times that a crop name appears in PHB-KG *(execution time* < 1*s)*.

```
SELECT ?cropName (count(?cropName) AS ?nb)
WHERE {
  SERVICE <http://ontology.inrae.fr/frenchcropusage/sparql> { GRAPH fcu:3.1 {
    ?body a skos:Concept ; skos:prefLabel ?cropName . FILTER (LANG(?cropName)='fr') }}
  ?phb a d2kab_inrae:Bulletin ; dct:spatial wikidata:Q16994 ; dul:isRealizedBy ?phbHtml .
  ?phbHtml dce:format "text/html" .
  ?rs a oa:ResourceSelection ; oa:hasSource ?phbHtml .
  ?aa a oa:Annotation ; oa:hasTarget ?rs ; oa:hasBody ?body .
} GROUP BY ?cropName ORDER BY DESC(?nb)
```

Listing 9: SPARQL query implementing CQ7 and retrieving the number of times that each crop name is mentioned in PHB bulletins related to the '*Pays de la Loire*' French region *(execution time* ≈ 35*s)*.

Combined exploitation of WheatGenomicsSLKG and RiceGenomicsSLKG CQ5 requires to use both wheat and rice KGs to search a correlation between gene expression and disease resistance. The aim is to search for wheat and rice genes co-occurring with the same taxon of a pathogen (or a more specific taxon) to identify candidate orthologous genes in wheat and rice genomes. The SPARQL query presented in Listing 12 implements this competency question. Starting with the *Magnaporthe oryzae* URI[38] or its upper parent,[39] we retrieve wheat and rice genes co-occurring with these taxa. This may be the indication of orthologous genes in wheat and rice genomes that should be explored by agronomists.

[38] http://purl.obolibrary.org/obo/NCBITaxon_318829.
[39] http://purl.obolibrary.org/obo/NCBITaxon_639021.

```
SELECT ?phbDescription ?cropName ?growthStageName ?xpath ?phbUri WHERE {
    SERVICE <http://ontology.inrae.fr/ppdo/sparql> {
    ?bodyDevt a owl:NamedIndividual ;
       skos:inScheme <http://ontology.inrae.fr/ppdo/ontology/bbch_globalScale> ;
       skos:prefLabel ?growthStageName . FILTER (LANG(?growthStageName)='fr') }
   ?phbUri a d2kab_inrae:Bulletin ; dul:isRealizedBy ?phbHtml ; dct:description ?phbDescription.
   ?phbHtml dce:format "text/html" .
   ?rs1 a oa:ResourceSelection ; oa:hasSource ?phbHtml ; oa:hasSelector ?sel1 .
   ?aa1 a oa:Annotation ; oa:hasTarget ?rs1 ; oa:hasBody ?bodyFcu .
   ?sel1 a oa:XPathSelector ; rdf:value ?xpt .
   ?rs2 a oa:ResourceSelection ; oa:hasSource ?phbHtml ; oa:hasSelector ?sel2 .
   ?sel2 a oa:XPathSelector ; rdf:value ?xpath .
   ?aa2 a oa:Annotation ; oa:hasTarget ?rs2 ; oa:hasBody ?bodyDevt .
   SERVICE <http://ontology.inrae.fr/frenchcropusage/sparql> { GRAPH fcu:3.1 {
     ?bodyFcu a skos:Concept ; skos:prefLabel ?cropName . FILTER (LANG(?cropName)='fr') }}
} LIMIT 20
```

Listing 10: SPARQL query implementing CQ8 and retrieving couples of annotations related respectively to crop names and development stages appearing in the same HTML element of a PHB *(execution time* $\approx 2s$*)*.

```
SELECT ?phbDescription ?positionCropName ?cropName ?positionGrowthStageName ?growthStageName
WHERE {
  SERVICE <http://ontology.inrae.fr/ppdo/sparql> {
    ?bodyDevt a owl:NamedIndividual ;
       skos:inScheme <http://ontology.inrae.fr/ppdo/ontology/bbch_globalScale> ;
       skos:prefLabel ?growthStageName . FILTER (LANG(?growthStageName)='fr') }
  ?phb a d2kab_inrae:Bulletin ; dul:isRealizedBy ?phbHtml ; dct:description ?phbDescription .
  ?phbHtml dce:format "text/html" .
  ?rs1 a oa:ResourceSelection ; oa:hasSource ?phbHtml ; oa:hasSelector ?sel1 .
  ?aa1 a oa:Annotation ; oa:hasTarget ?rs1 ; oa:hasBody ?bodyFcu .
  ?sel1 a oa:TextPositionSelector ; oa:start ?positionCropName .
  ?rs2 a oa:ResourceSelection ; oa:hasSource ?phbHtml ; oa:hasSelector ?sel2 .
  ?sel2 a oa:TextPositionSelector ; oa:start ?positionGrowthStageName .
  FILTER (abs(?positionGrowthStageName-?positionCropName) < 1000)
  ?aa2 a oa:Annotation ; oa:hasTarget ?rs2 ; oa:hasBody ?bodyDevt .
  SERVICE <http://ontology.inrae.fr/frenchcropusage/sparql> { GRAPH fcu:3.1 {
    ?bodyFcu a skos:Concept ; skos:prefLabel ?cropName . FILTER (LANG(?cropName)='fr') }}
} LIMIT 20
```

Listing 11: Another SPARQL implementation of CQ8 retrieving couples of annotations located within a window of a fixed number of characters *(execution time* $\approx 3s$*)*.

```
SELECT distinct ?paper ?title (GROUP_CONCAT(distinct ?geneName; SEPARATOR="-") as ?genes)
  ?ncbiTaxon
WHERE {
  ?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 ?ncbitaxonURI .
  ?source2 frbr:partOf+ ?paper .
  ?paper a fabio:ResearchPaper ; dct:title ?titleURI .
  ?titleURI rdf:value ?title .
  GRAPH <http://purl.obolibrary.org/obo/ncbitaxon/ncbitaxon.owl> {
    ?ncbitaxonURI rdfs:subClassOf* <http://purl.obolibrary.org/obo/NCBITaxon_639021> ;
      rdfs:label ?ncbiTaxon . }
}
```

Listing 12: SPARQL federated query implementing CQ5 and allowing the combined exploitation of WheatGenomicsSLKG and RiceGenomicsSLKG to retrieve orthologous genes (execution time ≈ 0.95 s).

Combined exploitation of WheatGenomicsSLKG and PHB KG The SPARQL query presented in Listing 13 implements CQ9: it enables to retrieve publications in PubMed and PHB bulletins corpora mentioning the same taxon (*Triticum aestivum*). As each corpus uses different semantic resources to annotate taxon entities (NCBI taxonomy in WheatGenomicsSLKG, and FCU thesaurus in PHB KG), the query exploits a third KG, TAXREF-LD[40] [13] to retrieve the alignments between NCBI classes and FCU concepts (line 20). Alignments between FCU concepts and TAXREF-

[40] TAXREF-LD is an RDF knowledge graph representing TAXREF, the French national taxonomical register for fauna, flora and fungus. Documentation of TAXREF-LD is available at https://github.com/frmichel/taxref-ld.

```
SELECT distinct ?paper ?bsv ?taxLabel ?fcuCropName ?taxrefClass
FROM <http://ns.inria.fr/d2kab/graph/wheatgenomicsslkg>
FROM <http://ns.inria.fr/d2kab/graph/alignments-fcu-taxref>
WHERE {
  { SELECT distinct ?paper ?taxon WHERE {
      ?annot a oa:Annotation; oa:hasTarget [ oa:hasSource ?source ] ; oa:hasBody ?taxon .
      ?taxon a d2kab:Taxon; skos:prefLabel ?label .
      ?source frbr:partOf+ ?paper .
      ?paper a fabio:ResearchPaper ; dct:title ?source .
      FILTER(CONTAINS(?label, "Triticum aestivum"))
    } LIMIT 100 }
  SERVICE <http://taxref.i3s.unice.fr/sparql> {
    ?taxrefClass owl:equivalentClass ?taxon ; rdfs:label ?taxLabel . }
  ?fcuCropName taxref:candidateAlignment_eppo|taxref:candidateAlignment_geves ?taxrefClass .
  SERVICE <http://ontology.inrae.fr/bsv/sparql> {
    ?bsv a d2kab_inrae:Bulletin; dul:isRealizedBy ?s ; dct:spatial ?w ; dct:date ?date_bsv .
    ?aa a oa:Annotation ; oa:hasTarget [ oa:hasSource ?s ] ; oa:hasBody ?fcuCropName . }
} LIMIT 20
```

Listing 13: SPARQL federated query implementing CQ9 and allowing the combined exploitation of WheatGenomicsSLKG and PHB KGs (execution time ≈ 21.18 s).

LD classes were generated automatically based on the *Official Catalogue of Species and Varieties of Cultivated Crops*, which is denoted by the alignment predicate `taxref:candidateAlignment_geves`. This way, the query retrieves that taxon http://taxref.mnhn.fr/lod/taxon/127692 is aligned with FCU concepts `fcu:Bles_tendres_hiver`[41] and `fcu:Bles_tendres_printemps`.[42] Note that this example query is meant to be executed on the WheatGenomicsSLKG SPARQL endpoint, invoking the two other SPARQL endpoints via SERVICE clauses. This illustrates the fact that publishing KGs according to FAIR design and publication principles allows to achieve the important goal of querying uniformly several interoperable knowledge graphs.

5. Discussion and lessons learned

We produced three different knowledge graphs compliant with the semantic model presented in Section 3.2 and representing the annotations of three text corpora with automatically extracted NEs. The proposed semantic model is generic and independent from the NLP tools used for NERL and may be reused for representing annotations extracted using any NLP tools.

The annotations extracted from three corpora rely on the same understanding of the OA model. Each annotation links one mention (e.g., surface form) in a document to one domain-specific entity defined in external semantic resources. However, as illustrated in Fig. 6, two different annotations may annotate the same piece of text, hence sharing the same target resource (an entity mention in the PHB document) while having two different bodies (two different domain concepts).

As presented in Section 3.2, the OA model provides different types of selectors which are generic enough to fulfill different needs considering the representation of the resource source. To represent the annotations of the PubMed and PHB corpora, we used three different selectors proposed by OA to precisely locate an entity mention in a text: `oa:TextQuoteSelector`, `oa:TextPositionSelector`, and `oa:XPathSelector`. The first two selectors are used in both corpora. In particular the `oa:TextPositionSelector` selector locates the entity mention within this part of the document. `oa:XPathSelector` is used for the PHB corpus to identify the HTML element in which an entity mention was recognized. Note that the document structures are not initially represented in the same way in both corpora. In the PubMed corpus, each document is identified by a URI which corresponds to its entry in the PubMed repository. Entity mentions may be identified in the title, abstract or abstract's sub-parts, each having its own URI and being linked by the `frbr:partOf` property (Fig. 7). These URI are used to identify the source of OA annotations while the *oa:TextPositionSelector* selector locates them within it. In the PHB corpus, an HTML document is identified by a URI that is the source of the annotation. This work illustrates that OA selectors are sufficiently broad to locate entities mentions in a variety of situations. Furthermore, the OA model can identify closely related annotations within a text, whether defined by a specific number of characters or by inclusion in the same structural element. This allows us to identify close entities in the text that may potentially be linked by a semantic relationship, as expressed by several competency questions. The coverage scope of our model could be extended by identifying complementary vocabularies. In particular, information such as frequency, lexical and morphological characteristics that can be drawn from text corpora and semantic resources can be added to our model by reusing terms from the FrAC vocabulary, an OntoLex module for Frequency, Attestation and Corpus information (FrAC) [3]. FrAC allows to model absolute frequencies of a given lexical entity (how many times an element of a semantic resource is recognized in the text, e.g., as shown in CQ5 bis) which is a recurrent need. Fig. 11 presents an RDF graph that links six instances of class `oa:Annotations` to one instance of the `frac:CorpusFrequency` class using the `prov:wasDerivedFrom` property. Those annotations are about the FCU concept "grapevine" recognized in the PHB example of Fig. 2. Considering that a document is a corpus of one element, the instance of class `frac:CorpusFrequency` represents the number of times that the given concept (grapevine) is recognized in the PHB document.

Regarding the effectiveness of NLP processes, we noticed that it largely depends on the quality and coverage of the semantic resources used. We have observed a lexical gap between domain-specific terms found in text corpora and those in the available semantic resources. For example, the FCU thesaurus adequately covers crop names, encompassing a broad range of terms found in PHB texts. Note that we have previously updated the FCU thesaurus to align it with the crop labels utilized in PHB documents as presented in [2]. However, growth stages present more variability in their expression, which poses challenges for existing semantic resources. This variability often results in instances not being identified by the pipeline. For instance, the term "pré-floraison" (pre-flowering) was overlooked as a growth stage because its surface form did not correspond to any of the entries associated with this stage, such as "boutons floraux séparés" (separated flowering buds) or "ifv label 57" (a code within the IFV label's scale).

The different NLP tools used for each dataset show that the KG creation methodology that we advance is independent from the NLP process. Indeed the automatic annotation of texts can be improved or

[41] https://ontology.inrae.fr/frenchcropusage/res/Bles_tendres_hiver.
[42] https://ontology.inrae.fr/frenchcropusage/page/res/Bles_durs_printemps.

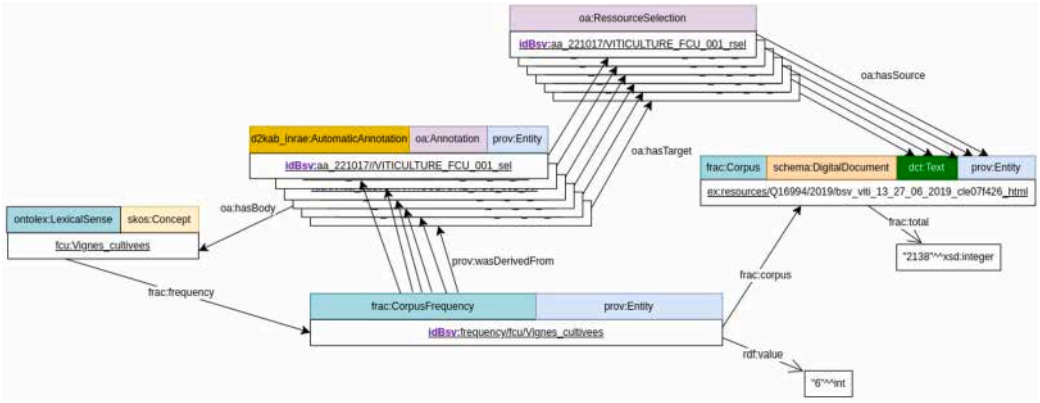


Fig. 11. Frequency Modelling in PHB graph based on FRAC, OA and PROV vocabularies.

replaced with another tool to improve either the accuracy or the scalability.

6. Conclusions and future works

In this paper, we presented the results of a research work to support scientists with methodologies to standardize and share domain knowledge extracted from texts according to FAIR principles. We rely on Semantic Web models and technologies to build domain-specific KGs that allow agronomists to explore and retrieve information from the annotated corpora and deduce new domain knowledge. Our approach relies on the formalization of a unified semantic data model to describe, structure and integrate annotations using NE automatically recognized from texts in the agricultural domain. NE entities normalization and linking is based on semantic resources widely adopted in Agriculture (for phenotypes, traits, taxa, cultivated varieties). The core part of this model is based on the W3C Web Annotation Ontology (OA) which has been complemented by eight different vocabularies to describe documents metadata and provenance information. We used this model to construct three different knowledge graphs from three distinct agricultural corpora using a mapping-based transformation pipeline. The relevance of the semantic model was validated by implementing a set of competency questions with SPARQL queries which reflect how the KGs can be queried to retrieve co-occurrence of NE in texts. The proposed semantic model and generation pipeline are generic enough to be reused to build new KGs in different research domains in order to enable scientists explore their scientific literature.

As future works, we want to assess then improve the quality of the extraction process. In particular we need to investigate the extraction of relations between recognized NE. Several competency questions involve retrieving entities that appear in the same context within texts. They could be refined by precising the relationship between the entities. As relation extraction strongly depends on the accuracy of the entity recognition task, an important first step for the PHB knowledge graph will focus on the improvement of the accuracy of NE annotations which is not always satisfying so far. Indeed, the evaluation and the use of state-of-the art supervised and fine-tuning models require training and benchmarking data, which to our knowledge is seldom available in the field of plant sciences. We plan to annotate and publish gold standard datasets based on the three corpora. This will require considerable efforts of domain experts to define guidelines and samples for NE and relation annotations. Gold-standard datasets can be used to train and evaluate natural language processing (NLP) approaches, such as specialized NE recognition, relation extraction and entity linking. As part of future work, we also plan to address various issues related to the entire lifecycle of KG. Firstly, we aim to develop visualization services to enable agronomists, who may not have expertise in semantic web technologies, to navigate and explore the different graphs with ease. Additionally, we intend to deploy our methodology pipeline to facilitate systematic incremental updates of the KGs with new documents and entities.

Funding

This work was carried out within the project D2KAB “From Data to Knowledge in Agronomy and Biodiversity” financed by the French National Research Agency (ANR-18-CE23-0017).

Material availability

The SKOS version of the WTO used in this study can be queried: http://d2kab.i3s.unice.fr/sparql.

The NCBITaxon used for this study can be found in the OBO Foundry repository: https://obofoundry.org/ontology/ncbitaxon.html.

The materials used in this study (AlvisNLP outputs, xR2RML mapping rules) to produce the wheat genomics literature KG are available in the GitHub repository: https://github.com/Wimmics/d2kab-wheatGenomicsLiterature-kg.

The materials used in this study (Hunflair outputs, xR2RML mapping rules) to produce the rice genomics literature KG are available in the GitHub repository: https://github.com/ANR-DIG-AI/RiceGenomicsSLKG.

The SPARQL endpoint of the KG on wheat and rice genomics scientific literature: http://d2kab.i3s.unice.fr/sparql.

The Plant Health Bulletin sub-corpora, associated code sources (AlvisNLP plans, xR2RML mapping rules, ...) used for this study are available in the *Corpus de Bulletins de Santé du Végétal* repository: https://forgemia.inra.fr/bsv/corpus-bsv.

The version 3.2 of the FCU thesaurus used for this study can be found in the AgroPortal repository: https://agroportal.lirmm.fr/ontologies/CROPUSAGE.

The version 1.2 of PPDO used for this study can be found in the AgroPortal repository: https://agroportal.lirmm.fr/ontologies/PPDO.

The SPARQL endpoint of the PHB KG: http://ontology.inrae.fr/bsv/sparql.

CRediT authorship contribution statement

Nadia Yacoubi Ayadi: Writing – review & editing, Writing – original draft, Validation, Software, Resources, Methodology, Conceptualization. **Stephan Bernard:** Validation, Software, Resources. **Robert Bossy:** Validation, Supervision, Resources, Data curation. **Marine Courtin:** Data curation. **Bill Gates Happi Happi:** Software, Resources. **Pierre Larmande:** Validation, Supervision, Software, Resources. **Franck Michel:** Writing – review & editing, Software, Resources, Conceptualization. **Claire Nédellec:** Writing – review & editing, Validation, Supervision. **Catherine Roussey:** Writing – original

draft, Validation, Resources. **Catherine Faron:** Writing – review & editing, Writing – original draft, Validation, Project administration, Methodology, Conceptualization.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Catherine Faron reports financial support was provided by French National Research Agency. Claire Nedellec reports financial support was provided by French National Research Agency. Catherine Roussey reports financial support was provided by French National Research Agency. Pierre Larmande reports financial support was provided by French National Research Agency. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data will be made available on request.

References

[1] Mouhamadou Ba, 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) Organised with LREC 2016, Portorož, Slovenia, 2016, pp. 15–18.

[2] Anna Chepaikina, Robert Bossy, Catherine Roussey, Stephan Bernard, Thesaurus enrichment via coordination extraction, in: 16th International Conference on Metadata and Semantics Research (MTSR 2022), London, United Kingdom, November 2022.

[3] Christian Chiarcos, Maxim Ionov, Jesse de Does, Katrien Depuydt, Fahad Khan, Sander Stolk, Thierry Declerck, John Philip McCrae, Modelling frequency and attestations for ontolex-lemon, in: Proceedings of the 2020 Globalex Workshop on Linked Lexicography, 2020, pp. 1–9.

[4] Brett Drury, Robson Fernandes, Maria-Fernanda Moura, Alneu de Andrade Lopes, A survey of semantic web technology for agriculture, Inf. Process. Agric. 6 (4) (2019) 487–501.

[5] Clément Jonquet, Anne Toulet, Elizabeth Arnaud, Sophie Aubin, Esther Dzalé Yeumo, Vincent Emonet, John Graybeal, Marie-Angélique Laporte, Mark A. Musen, Valeria Pesce, Pierre Larmande, Agroportal: a vocabulary and ontology repository for agronomy, Comput. Electron. Agric. 144 (2018) 126–143.

[6] Anas Fahad Khan, Christian Chiarcos, Thierry Declerck, Daniela Gifu, Elena González-Blanco García, Jorge Gracia, Maxim Ionov, Penny Labropoulou, Francesco Mambrini, John P. McCrae, Émilie Pagé-Perron, Marco Passarotti, Salvador Ros Muñoz, Ciprian-Octavian Truica, When linguistics meets web technologies. Recent advances in modelling linguistic linked data, Semant. Web 13 (6) (2022) 987–1050.

[7] Kurata Nori, Yukiko Yamazaki, Oryzabase. An integrated biological and genome information database for rice, Plant Physiol. 140 (1) (January 2006) 12.

[8] F. Michel, L. Djimenou, C. Faron-Zucker, J. Montagnat, Translation of relational and non-relational databases into RDF with xR2RML, in: Proceeding of the 11th International Conference on Web Information Systems and Technologies (WebIST), Lisbon, Portugal, 2015, pp. 443–454.

[9] Uwe Meier, Growth Stages of Mono- and Dicotyledonous Plants: BBCH Monograph, Open Agrar Repositorium, 2018.

[10] Franck Michel, Catherine Faron-Zucker, Olivier Corby, Fabien Gandon, Enabling automatic discovery and querying of Web APIs at Web scale using linked data standards, in: Companion Proceedings of the World Wide Web Conference 2019 - WWW 19, San Francisco, USA, ACM Press, 2019, pp. 883–892.

[11] Luc Moreau, Paul Groth, Provenance: an Introduction to PROV, Synthesis Lectures on the Semantic Web: Theory and Technology, vol. 3(4), 2013, pp. 1–129.

[12] 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, in: Lecture Notes in Computer Science, vol. 12507, Springer, November 2020, pp. 294–310.

[13] Franck Michel, Olivier Gargominy, Sandrine Tercerie, Catherine Faron-Zucker, A model to represent nomenclatural and taxonomic information as linked data. Application to the French taxonomic register, TAXREF, in: Proceedings of the ISWC2017 Workshop on Semantics for Biodiversity (S4BioDiv), Vienna, Austria, in: CEUR Workshop Proceedings, vol. 1933, 2017, http://ceur-ws.org/Vol-1933/paper-3.pdf.

[14] Jose Martinez-Rodriguez, Aidan Hogan, Ivan Lopez-Arevalo, Information extraction meets the semantic web: a survey, Semant. Web 11 (2020) 255–335.

[15] Claire Nédellec, Robert Bossy, Dialekti Valsamou, Marion Ranoux, Wiktoria Golik, Pierre Sourdille, Information extraction from bibliography for marker-assisted selection in wheat, in: Proceedings of Research Conference on Metadata and Semantics Research - MTSR 2014, 2014, pp. 301–313.

[16] Claire Nédellec, Liliana L. Ibanescu, Robert Bossy, Pierre Sourdille, WTO, an ontology for wheat traits and phenotypes in scientific publications, Genomics Inform. 18 (2) (2020).

[17] Silvio Peroni, SAMOD: an agile methodology for the development of ontologies, in: Mauro Dragoni, Maréa Poveda-Villalón, Ernesto Jimenez-Ruiz (Eds.), OWL: Experiences and Directions – Reasoner Evaluation, Springer, 2016, pp. 55–69.

[18] Valentina Presutti, Aldo Gangemi, Content ontology design patterns as practical building blocks for web ontologies, in: International Conference on Conceptual Modeling, Springer, 2008, pp. 128–141.

[19] Silvio Peroni, David Shotton, FaBiO and CiTO: ontologies for describing bibliographic resources and citations, J. Web Semant. 17 (2012) 33–43.

[20] Catherine Roussey, Stephan Bernard, François Pinet, Xavier Reboud, Vincent Cellier, Ivan Sivadon, Danièle Simonneau, Anne-Laure Bourigault, A methodology for the publication of agricultural alert bulletins as lod, Comput. Electron. Agric. 142 (2017) 632–650.

[21] Catherine Roussey, Xavier Delpuech, Florence Amardeilh, Stephan Bernard, Clement Jonquet, Semantic description of plant phenological development stages, starting with grapevine, in: Emmanouel Garoufallou, María-Antonia Ovalle-Perandones (Eds.), Metadata and Semantic Research, Springer International Publishing, Cham, 2021, pp. 257–268.

[22] Conrad L. Schoch, Stacy Ciufo, Mikhail Domrachev, Carol L. Hotton, Sivakumar Kannan, Rogneda Khovanskaya, Detlef D. Leipe, Richard McVeigh, Kathleen O'Neill, Barbara Robbertse, Shobha Sharma, Vladimir Soussov, John P. Sullivan, Lu Sun, Seán Turner, Ilene Karsch-Mizrachi, Ncbi taxonomy: a comprehensive update on curation, resources and tools, Database J. Biol. Databases Curation (2020) 2020.

[23] Robert Sanderson, Paolo Ciccarese, Benjamin Young, Web annotation ontology, Technical report, W3C, 2017.

[24] The UniProt Consortium, UniProt: the Universal Protein Knowledgebase in 2023, Nucleic Acids Res. 51 (D1) (11 2022) D523–D531.

[25] Mike Uschold, Michael Gruninger, Ontologies: principles, methods and applications, Knowl. Eng. Rev. 11 (2) (1996) 93–136.

[26] Aravind Venkatesan, Gildas Tagny Ngompe, Nordine El Hassouni, Imene Chentli, Valentin Guignon, Clement Jonquet, Manuel Ruiz, Pierre Larmande, Agronomic Linked Data (AgroLD): a knowledge-based system to enable integrative biology in agronomy, PLoS ONE 13 (11) (2018) 1–17.

[27] Mark Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gaby Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Olavo Bonino da Silva Santos, Philip Bourne, Jildau Bouwman, Anthony Brookes, Tim Clark, Merce Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris Evelo, Richard Finkers, Barend Mons, The fair guiding principles for scientific data management and stewardship, Sci. Data 3 (03 2016).

[28] Leon Weber, Mario Sänger, Jannes Münchmeyer, Maryam Habibi, Ulf Leser, Alan Akbik, HunFlair: an easy-to-use tool for state-of-the-art biomedical named entity recognition, arXiv:2008.07347, August 2020.

[29] Eric Yao, Victoria C. Blake, Laurel Cooper, Charlene P. Wight, Steve Michel, H. Busra Cagirici, Gerard R. Lazo, Clay L. Birkett, David J. Waring, Jean-Luc Jannink, Ian Holmes, Amanda J. Waters, David P. Eickholt, Taner Z. Sen, GrainGenes: a data-rich repository for small grains genetics and genomics, Database 2022 (05 2022) baac034.