A Unified Approach to Publish Semantic Annotations of Agricultural Documents as Knowledge Graphs

Nadia Yacoubi Ayadi, Stephan Bernard, Robert Bossy, Marine Courtin, Bill Gates Happi Happi, Pierre Larmande, Franck Michel, Claire Nedellec, Catherine Roussey, Catherine Faron

A Unified Approach to Publish Semantic Annotations of Agricultural Documents as Knowledge Graphs

Nadia Yacoubi Ayadi[1,2], Stephan Bernard[3], Robert Bossy[4], Marine Courtin[3,4], Bill Gates Happi Happi[5], Pierre Larmande[5], Franck Michel[1], Claire Nedellec[4], Catherine Roussey[3,6], and Catherine Faron[1]

[1]Université Côte d'Azur, Inria, CNRS, I3S (UMR 7271), France

[2]Université de Lyon 1, CNRS, LIRIS (UMR 5205), France

[3]TSCF, INRAE, Centre Auvergne Rhône Alpes Clermont, Aubière, France

[4]MaIAGE, INRAE, Université Paris-Saclay, 78350 Jouy-en-Josas, France

[5]DIADE, IRD, CIRAD, Univ. Montpellier, Montpellier, France

[6]Mistea, INRAE, Centre Occitanie, Montpellier, France

January 31, 2024

Abstract

This paper presents a generic semantic model to describe, structure and integrate the named entities automatically extracted from texts coded as annotations. This model has been used to construct three different knowledge graphs from three distinct agricultural corpora in two different languages (English and French). The two first corpora consist of PubMed scientific publications, written in English, on wheat and rice genetics and phenotyping. The named entities to be recognised are genes, phenotypes, traits, genetic markers, and taxa. Those entities are normalized using domain semantic resources such as the Wheat Trait and Phenotype Ontology (WTO). The third corpus contains agricultural

alert bulletins published in France and written in French. The named entities to be recognised are crop names and phenological development stages. Crop names are defined in the French Crop Usage (FCU) thesaurus while development stages are formalized using the BBCH-based Plant Phenological Description Ontology (PPDO). For the three corpora, named entities were automatically extracted using natural language processing tools. We present an approach that relies on the formalization of a semantic data model based on common and well-adopted linked open vocabularies such as Web Annotation Ontology (OA) and Provenance ontology (PROV). The model describes the named entities and their links to vocabularies. It was slightly adapted to each corpus annotations. The model was populated using a mapping-based transformation pipeline implemented with the Morph-xR2RML tool which takes CSV files as input . The development of the proposed model was initiated by the formulation of motivating scenarios by experts in the domain of each corpus that led to a set of competency questions. Those questions provided requirements on the semantic model. The relevance of the semantic model was validated by implementing the competency questions into SPARQL queries enabling to query the constructed RDF knowledge graphs.

Keywords

Agriculture, Knowledge Graphs, Semantic modelling, RDF Transformation, Natural Language Processing, Annotations, Semantic Resources, Named Entity Recognition and Linking

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 [WDA+16] referring to best practice guidelines: resources must be accessible, understood, exchanged and reused by machines. 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 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 domain experts [KCD+22]. 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 a structured and meaningful annotation. 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. 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 paper we focus on building knowledge graphs representing semantic annotations of textual documents in the field of agriculture. 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. Two corpora are collections of scientific publications on rice and wheat functional genomics, retrieved from the PubMed[1] repository. These publications investigate the gene-phenotype link for varietal selection, or more precisely the identification of gene markers involved in the expression of a given phenotype, for selection assistance [NBV+14]. A third corpus gathers technical documents called Plant Health Bulletins (PHBs) which are agricultural alert bulletins published in France. Although, the KGs that we built from these corpora are intended to serve different needs, we have adopted the same methodology and core data model. The first step of our methodology leverages NLP pipelines to perform the tasks of Named Entity Recognition (NER) and Linking (NEL) [MRHLA20]. A semantic data model has been defined to capture how NE annotations

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

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 [Per16], which in turn is based on the early work of Uschold & Gruninger [GF95, UG96]. 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. All CQ demonstrate that the KGs should be uniformly queried by experts to highlight the context of NE co-occurrence in the original texts in order to reveal hidden interactions between NE. In the continuity of earlier works [MGA+20], we propose to rely on the Open Annotation Ontology (OA) [SCY17] 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 [MDFM15].

The paper is structured as follows. in Section 2, we present the materials of our research work which consist of three text corpora and the semantic resources used to annotate those corpora. in Section 3.1, we present the CQs of each case study. We describe the proposed semantic model in further details in Section 3.2. in Section 4, we present 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. For the linking task, NLP processes rely on a set of semantic resources presented in Section 2.2.

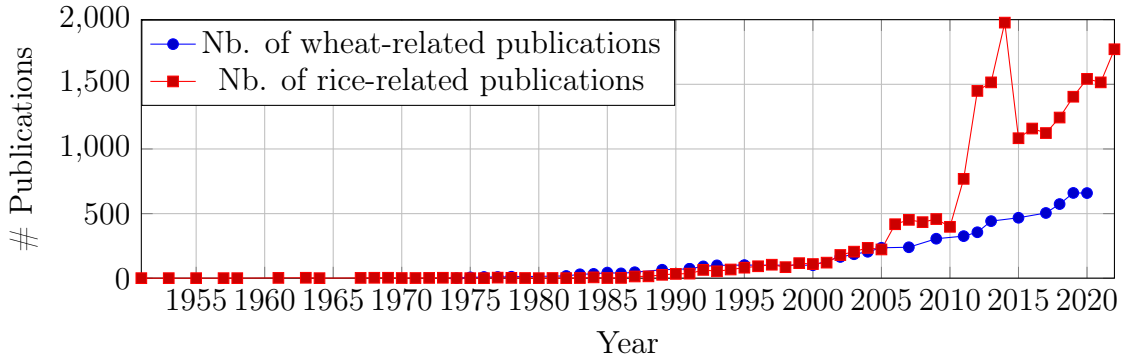


Figure 1: Evolution of the number of research works on wheat and rice genomics from 1951 to 2022

2.1 Text Corpora

2.1.1 PubMed Corpus on Wheat Genomics

The MaIAGE research group[2] has collected 8,496 scientific references of articles published between 1974 and 2021 related to ~~wheat genomics~~ wheat selection. A first exploration of the corpus shows that over the last two decades, the number of publications on ~~wheat genomics~~ has steadily increased. While the first publication appears in 1974, the annual publication grew slowly between 1975 and 2000. 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 Figure 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. We used the AlvisNLP pipeline to identify genes, traits, phenotypes, taxa and varieties entities mentioned in the title and the abstract of publications, and the relationships between wheat varieties and phenotypes. In total, 88,880 mentions of 4,318 distinct named entities were recognized and linked to exiting entities of semantic resources (presented in Section 2.2). Figure 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*), phenotypes (e.g., *leaf rust resistance*, *resistance to stem rust*, *powdery mildew resistance*) and taxa (e.g., *wheat*).

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

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

2.1.2 PubMed Corpus on Rice Genomics

The DIADE research group[3] collected 17,058 scientific articles from the Oryzabase database [KY06] 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 past two decades, which is consistent with the sharp increase in research interest in rice genetics as depicted in Figure 1. We used the HunFLAIR NER tagger [WSM+20] that we combined with NLTK, Spacy and other Python libraries to extract four types of named entities in the title or the abstract of the articles: we distinguish 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.

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 [RBP+17]. PHBs synthesize the interpretation of observations performed on crops by different collecting networks, elements from epidemiological models, meteorological data and sometimes

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

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 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 500 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, that is to say, the publication date may vary from 2009 to 2020. These three sub-corpora are available on a Git repository[4]. The whole corpus brings together a total of 880 PHBs with an average of 2,548 tokens per bulletin, covering the whole French territory and all crop categories of French agriculture. We use the AlvisNLP pipeline to extract NEs referring to french crop names and french development stages mentioned in the text of PHBs. In total, 37,488 mentions of 416 distinct NEs were extracted. Figure 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, thesaurii) was developed and published using Semantic Web technologies [DFMd19] and made available for research communities in open portals such the Agroportal repository[5] [JTA+18]. In this section, we present the semantic resources that we have reused to annotate the text corpora presented in Section 2.1.

[4] https://forgemia.inra.fr/bsv/corpus-bsv
[5] http://agroportal.lirmm.fr/

Figure 2: Example of expected NE recognition and linking in a grapevine PHB

2.2.1 Wheat Trait and Phenotype Ontology

The Wheat Trait and Phenotype Ontology (WTO) [NIBS20] 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[6] and contains 745 classes. We revised and transformed WTO 3.0 in OWL/SKOS format[7] and we used it to annotate mentions of phenotypes and traits recognised in the PubMed corpus on wheat functional genomics.

2.2.2 NCBI Taxonomy

The NCBITaxon ontology[8] 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 [SCD+20]. 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.

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

[7] https://github.com/Wimmics/d2kab-wheatGenomicsLiterature-kg/tree/main/wheat-kg-dataset/WTO-v3.0

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

2.2.3 French Crop Usage Thesaurus

The French Crop Usage (FCU) thesaurus[9] 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 Figure 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.

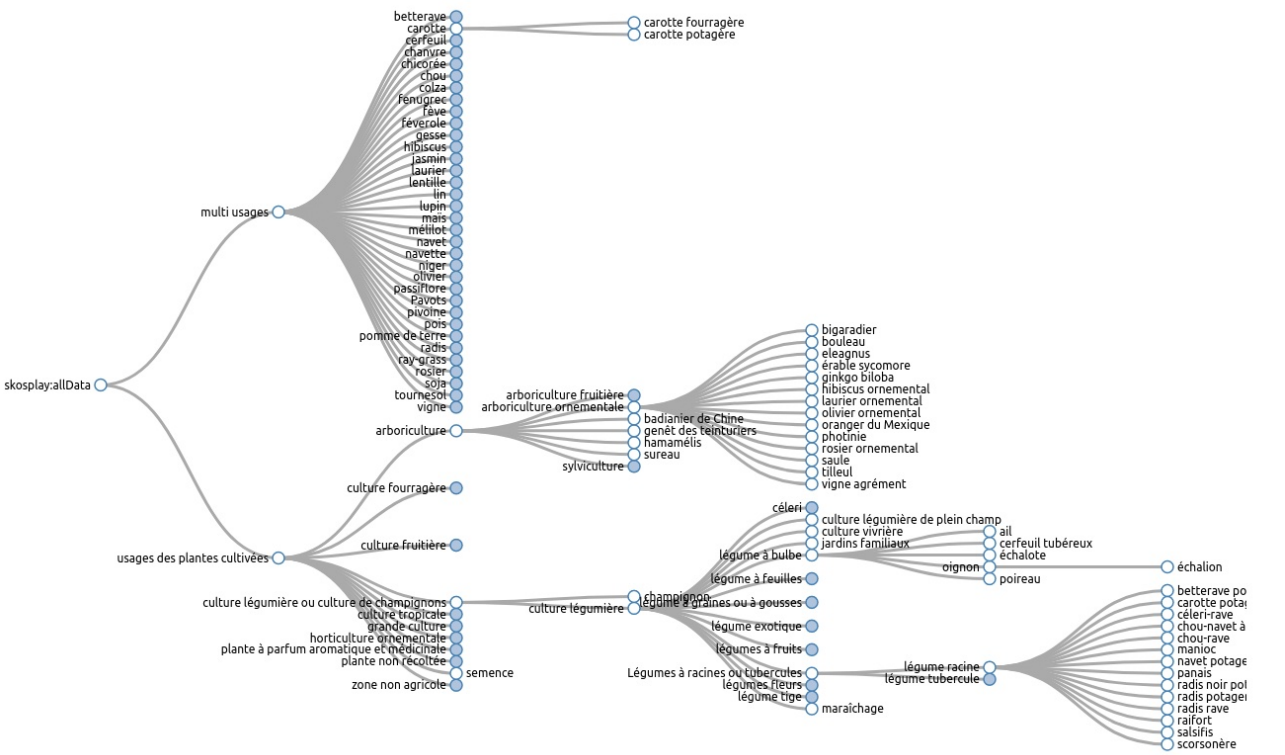


Figure 3: An extract from the FCU thesaurus. Visualisation generated by the SKOS Play tool.

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

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

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 [Mei18]. BBCH framework uniformly codes phenologically similar development stages of different plant species. The BBCH-based Plant Phenological Description Ontology (PPDO)[10] [RDA+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.

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 [The22] (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[11]. 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 [YBC+22] database was created. For rice genes, the Oryzabase [KY06] database was used and integrated into the AgroLD Knowledge Graph [VTNH+18] 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[12].

For wheat varieties recognition and normalization, a curated list was created combining two sources: (1) the *Plant variety catalogues, databases & information systems*[13] and (2) the

[10] The version 1.2 is available at https://agroportal.lirmm.fr/ontologies/PPDO
[11] https://sparql.uniprot.org/sparql/
[12] http://sparql.southgreen.fr
[13] https://food.ec.europa.eu/plants/plant-reproductive-material/plant-variety-catalogues-databases-information-systems

Official Catalogue of Species and Varieties of Cultivated Crops[14]. To be compliant with LOD principles, we created a URI to identify each distinct entry in the different created lexicons.

3 Method

3.1 Competency Questions

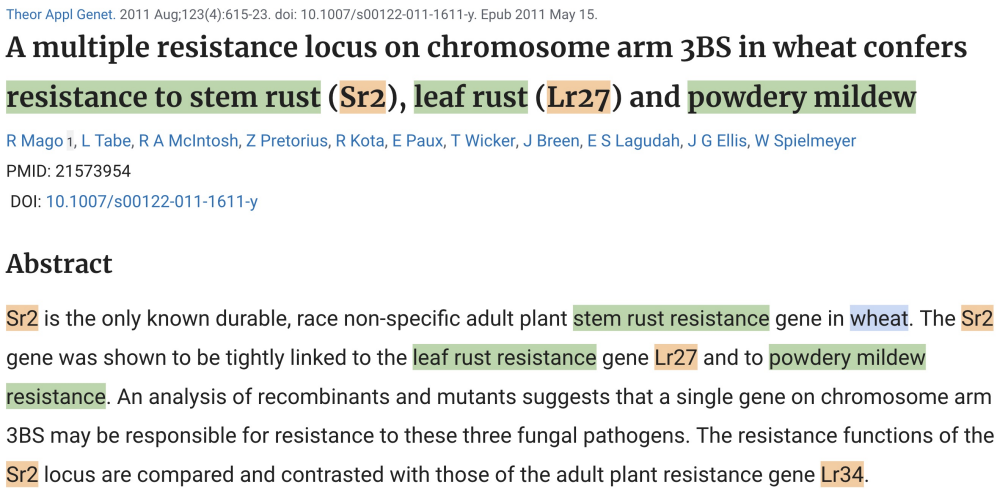
In this section, we present a set of Competency Questions (CQs) stemming from requirements expressed by experts and collected in the context of the D2KAB project[15]. CQs are natural language questions illustrating the typical knowledge that scientists would require a data source to provide. A common 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 (Figure 1) 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

[14] https://www.geves.fr/catalogue/
[15] https://www.d2kab.org/



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

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

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

PMID: 21573954

DOI: 10.1007/s00122-011-1611-y

Abstract

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

Figure 4: Example of NE recognition and linking in a PubMed publication

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 Figure 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 [NBV$^+$14]. 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 be involved in 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-occuring in literature with the same taxon of a pathogen (or a more specific taxon). This enables to identify orthologous genes[16] 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.

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

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 Figure 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 Figure 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 is related to a crop cultivated in the French territory, to search for example new crop varieties that are resistant to drought or high temperatures.

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 [MGA+20]. It is based on the W3C Web Annotation Ontology (OA) [SCY17] 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[17], "*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 structure of the text. 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.

Figure 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[18]. 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

[17] https://www.w3.org/TR/annotation-model/
[18] https://pubmed.ncbi.nlm.nih.gov/21573954/

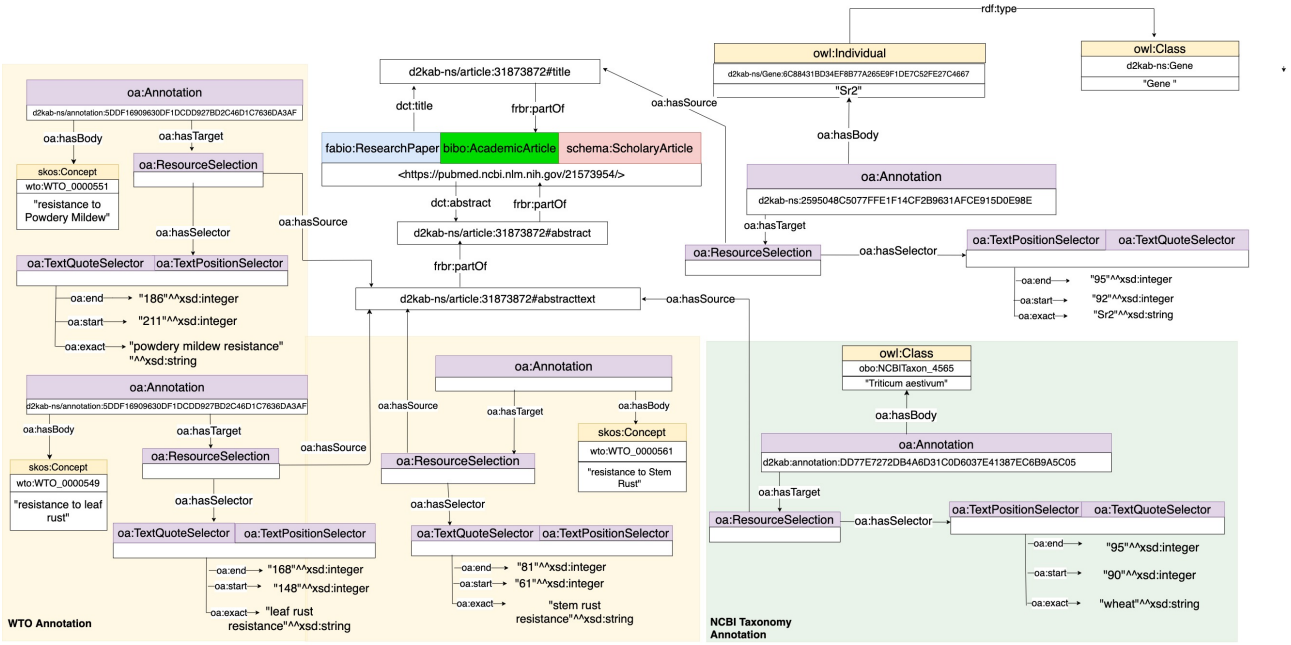


Figure 5: Example of NE annotations identified in a PubMed Publication's section (title and abstract) and represented in WheatGenomicsSLKG based on OA ontology.

in Figure 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 Figure 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.

Figure 6 represents three annotations extracted from the PHB[19] presented in Figure 2. One annotation[20] 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.

[19] http://ontology.inrae.fr/bsv/resources/Q16994/2019/bsv_viti_13_27_06_2019_cle07f426_html

[20] http://ontology.inrae.fr/bsv/resources/Q16994/2019/bsv_viti_13_27_06_2019_cle07f426/aa_221017/Vignes_cultivees_FCU_001

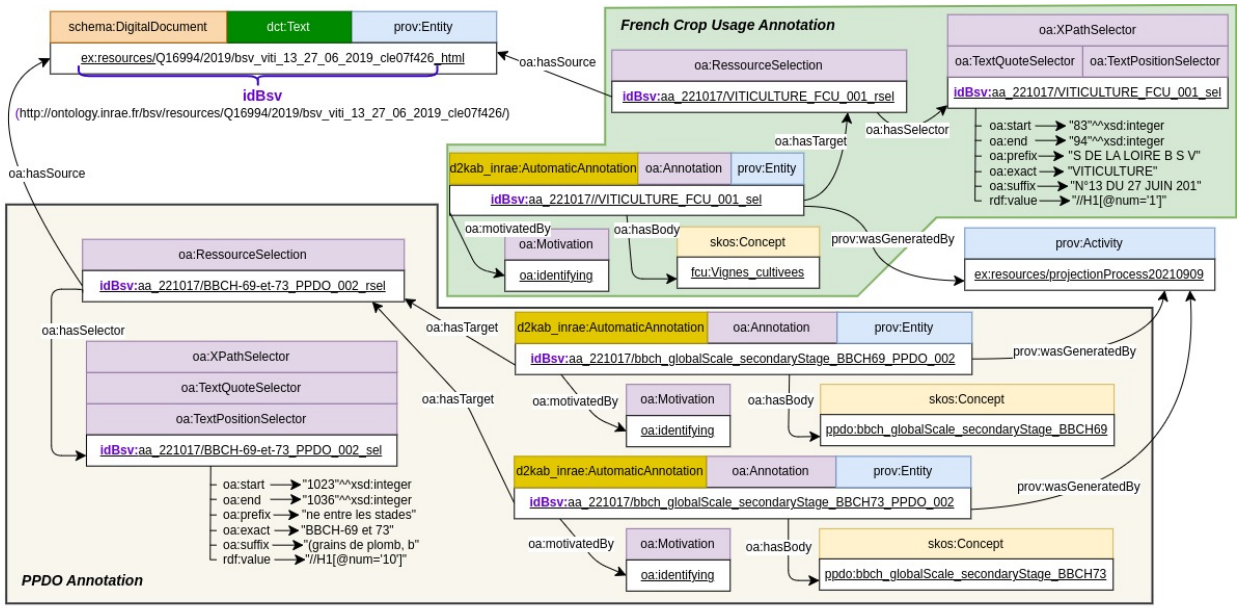


Figure 6: Example of annotations for a PHB. One annotation identifies crop, and the other one identifies two development stages.

- 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 Figure 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 aims 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 [21], FRBR aligned bibliographic ontology (FaBiO) [PS12], bibliographic ontology (BIBO)[22], Dolce Ultra Light (DUL) [PG08], PROV Ontology (PROV)

[21] https://www.dublincore.org/specifications/dublin-core/dcmi-terms/
[22] https://github.com/structureddynamics/Bibliographic-Ontology-BIBO

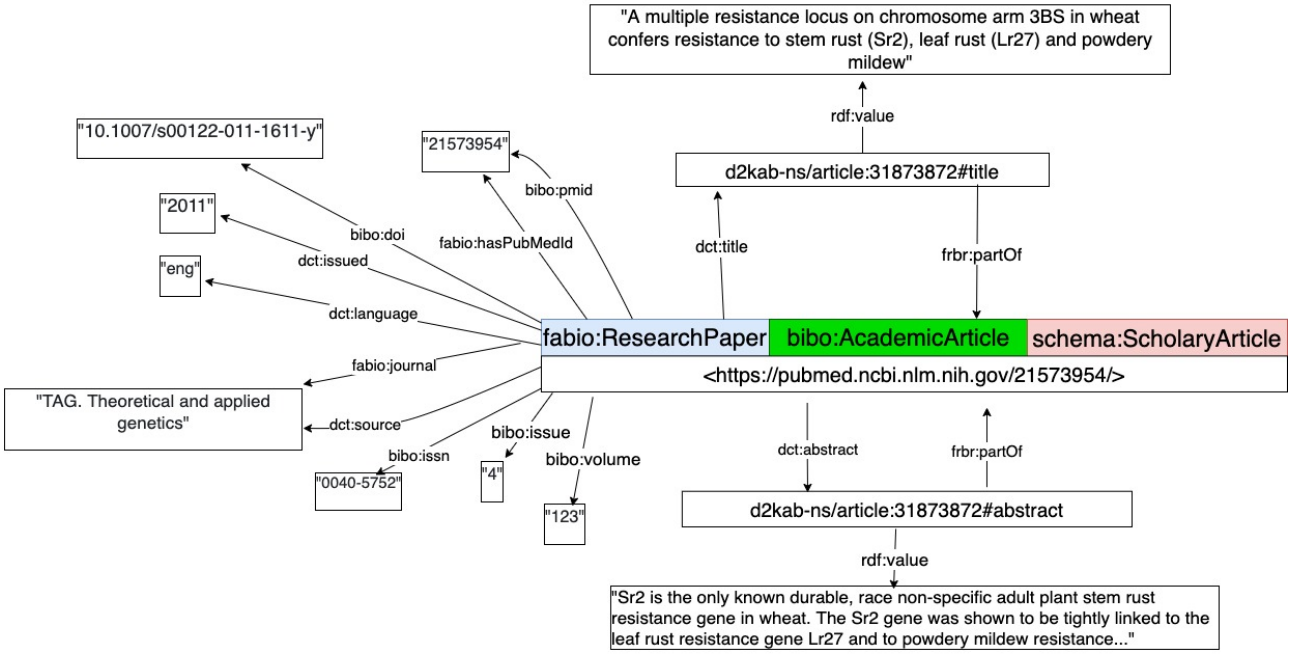


Figure 7: Example RDF graph describing bibliographic metadata of a scientific publication in the PubMed corpus

[MG13], the module for FRequency, Attestation and Corpus information (FRAC) of the LExicon Model for ONtologies (LEMON) [CIdD+20] 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 structured 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. Figure 7 illustrates (a subset of) the bibliographic metadata of a scientific document.

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#

Table 1: List of reused vocabularies

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` [PG08]. 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`.

Figure 8 presents the graph annotating the bulletin of Figure 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

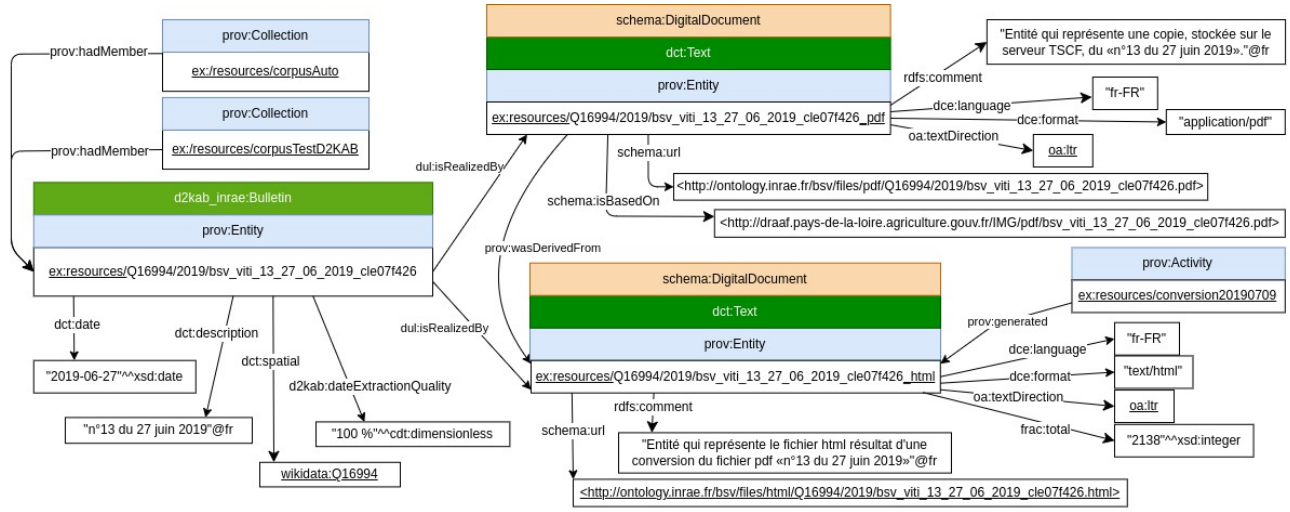


Figure 8: Example RDF graph describing metadata of a bulletin in the PHB corpus.

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 Figure 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.

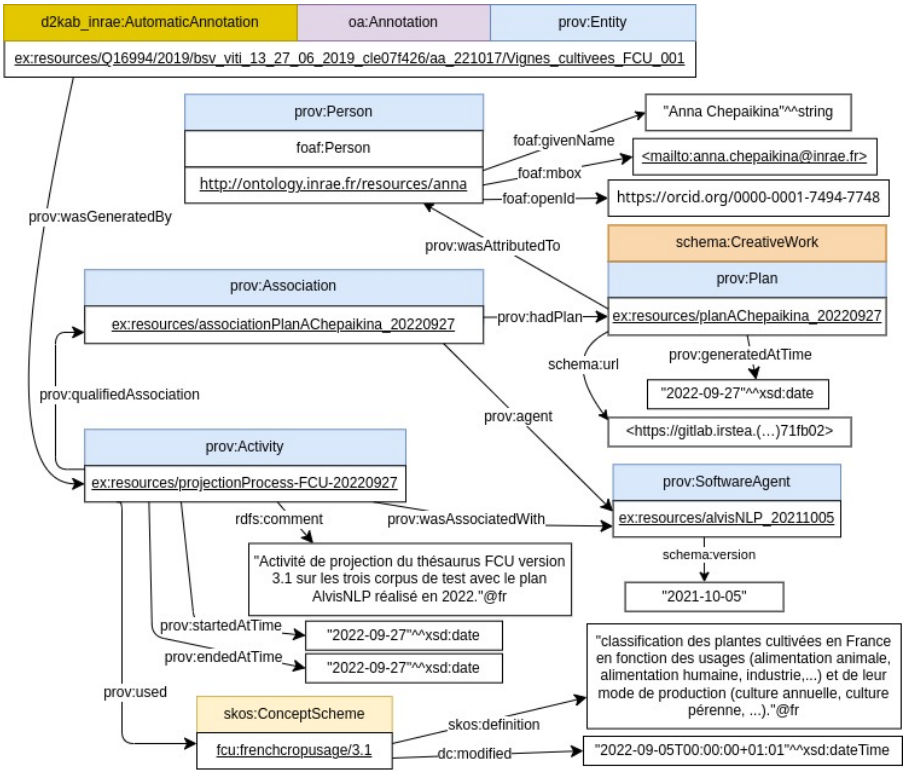


Figure 9: Example RDF graph describing provenance metadata of a bulletin in the PHB corpus.

4 Results and Validation

4.1 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 [MDFM15] 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[23]. 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.

4.1.1 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

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

	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}

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

directory[24]. Similar 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[25]. 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 [MFZCG19] to query the PubMed Central API and retrieve publication metadata[26]. 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. WheatGenomicsSLKG and RiceGenomicsSLKG can be queried at http://d2kab.i3s.unice.fr/sparql. A list of SPARQL queries which implement CQs (CQ1 to CQ5) are presented in Section 4.2.

4.1.2 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 are extracted (download date, download URL, local filename and web path) [RDA+21]. These data are transformed into RDF using python scripts.

[24] https://github.com/Wimmics/d2kab-wheatGenomicsLiterature-kg/tree/main/mapping-rules
[25] https://github.com/ANR-DIG-AI/RiceGenomicsSLKG
[26] https://sparql-micro-services.org/service/pubmed/getArticleByPMId_sd/

The downloaded pdf files are transformed into HTML using the `pdf2blocks`[27] 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[28]. The resulting RDF KG can be queried at http://ontology.inrae.fr/bsv/sparql. A list of SPARQL queries which implement CQs (CQ6 to CQ10) are presented in Section 4.2.

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 are quiet similar.

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[29]. 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 [30].

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 confirms that Lr34 is the most cited gene in the literature. Lr10, Lr26 and Lr24 genes appear

[27] https://doi.org/10.5281/zenodo.4067965
[28] https://forgemia.inra.fr/stephan.bernard/corpus-bsv
[29] http://d2kab.i3s.unice.fr/sparql
[30] https://github.com/Wimmics/d2kab-wheatGenomicsLiterature-kg/blob/main/SPARQLQueries-JupyterNotebook.ipynb

```
SELECT ?GeneName (count(distinct ?paper) as ?NbOcc)
FROM NAMED <http://ns.inria.fr/d2kab/graph/wheatgenomicsslkg>
FROM NAMED <http://ns.inria.fr/d2kab/ontology/wto/v3>
WHERE {
  GRAPH <http://ns.inria.fr/d2kab/graph/wheatgenomicsslkg> {
    ?a1 a oa:Annotation;
        oa:hasTarget [ oa:hasSource ?source1 ] ;
        oa:hasBody ?WTOtraitURI .
    ?source1 frbr:partOf+ ?paper .
    ?a a oa:Annotation ;
      oa:hasTarget [ oa:hasSource ?source ] ;
      oa:hasBody [ a d2kab:Gene; skos:prefLabel ?GeneName ] .
    ?source frbr:partOf+ ?paper .
    ?paper a fabio:ResearchPaper . }
  GRAPH <http://ns.inria.fr/d2kab/ontology/wto/v3> {
    ?WTOtraitURI skos:prefLabel "resistance to Leaf rust" . }
}
GROUP BY ?GeneName
HAVING (count(distinct ?paper) > 1)
ORDER BY DESC(?NbOcc)
```

Listing 1: SPARQL query implementing CQ1 and retrieving the most cited genes mentioned proximal to the "*resistance to Leaf rust*" trait in WheatGenomicsSLKG.

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 returns a list of scientific publications from the WheatGenomicsSLKG graph that list several genetic markers and genes entities 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

```
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*

```
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

```
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.

```
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.

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*[31]).

[31] http://purl.obolibrary.org/obo/NCBITaxon_5296

```
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).

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 [32].

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 (Figure 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

[32] https://forgemia.inra.fr/bsv/corpus-bsv/-/blob/SAAD/sample/PHB-KG_SPARQL_Queries.ipynb

```
SELECT ?phb ?cropName ?xpath
WHERE {
  SERVICE <http://ontology.inrae.fr/frenchcropusage/sparql> {
    ?body a skos:Concept ; skos:prefLabel ?cropName . FILTER (LANG(?cropName)='fr') . }
  ?phb a d2kab_inrae:Bulletin ; dul:isRealizedBy ?phb_html .
  ?phb_html dce:format "text/html" .
  ?rs a oa:ResourceSelection ; oa:hasSelector ?sel ; oa:hasSource ?phb_html .
  ?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

```
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 ?phb_html .
  ?phb_html dce:format "text/html" .
  ?rs a oa:ResourceSelection ; oa:hasSource ?phb_html .
  ?aa a oa:Annotation ; oa:hasTarget ?rs ; oa:hasBody ?body .
  SERVICE <http://ontology.inrae.fr/frenchcropusage/sparql> {
    ?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

of times that each crop name is mentioned in the 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 [RBP+17] 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

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

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 (50s) due to the call of two service templates.

```
SELECT ?phb ?cropName ?devtName ?xpt
WHERE {
  SERVICE <http://ontology.inrae.fr/ppdo/sparql> {
    ?body_devt a owl:NamedIndividual ;
      skos:inScheme <http://ontology.inrae.fr/ppdo/ontology/bbch_globalScale> ;
      skos:prefLabel ?devtName . FILTER (LANG(?devtName)='fr') }
  ?phb a d2kab_inrae:Bulletin ; dul:isRealizedBy ?phb_html .
  ?phb_html dce:format "text/html" .
  ?rs1 a oa:ResourceSelection ; oa:hasSource ?phb_html ; oa:hasSelector ?sel1 .
  ?aa1 a oa:Annotation ; oa:hasTarget ?rs1 ; oa:hasBody ?body_fcu .
  ?sel1 a oa:XPathSelector ; rdf:value ?xpt .
  ?rs2 a oa:ResourceSelection ; oa:hasSource ?phb_html ; oa:hasSelector ?sel2 .
  ?sel2 a oa:XPathSelector ; rdf:value ?xpt .
  ?aa2 a oa:Annotation ; oa:hasTarget ?rs2 ; oa:hasBody ?body_devt .
  SERVICE <http://ontology.inrae.fr/frenchcropusage/sparql> {
    ?body_fcu a skos:Concept ; skos:prefLabel ?cropName . FILTER (LANG(?cropName)='fr') }
} LIMIT 10
```

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.

In the PHB example of Figure 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.

```
SELECT ?phb ?pos1 ?cropName ?pos2 ?devtName
WHERE {
  SERVICE <http://ontology.inrae.fr/ppdo/sparql> {
    ?body_devt a owl:NamedIndividual ;
       skos:inScheme <http://ontology.inrae.fr/ppdo/ontology/bbch_globalScale> ;
       skos:prefLabel ?devtName . FILTER (LANG(?devtName)='fr') }
  ?phb a d2kab_inrae:Bulletin ; dul:isRealizedBy ?phb_html .
  ?phb_html dce:format "text/html" .
  ?rs1 a oa:ResourceSelection ; oa:hasSource ?phb_html ; oa:hasSelector ?sel1 .
  ?aa1 a oa:Annotation ; oa:hasTarget ?rs1 ; oa:hasBody ?body_fcu .
  ?sel1 a oa:TextPositionSelector ; oa:start ?pos1 .
  ?rs2 a oa:ResourceSelection ; oa:hasSource ?phb_html ; oa:hasSelector ?sel2 .
  ?sel2 a oa:TextPositionSelector ; oa:start ?pos2 .
  FILTER (abs(?pos2-?pos1) < 1000)
  ?aa2 a oa:Annotation ; oa:hasTarget ?rs2 ; oa:hasBody ?body_devt .
  SERVICE <http://ontology.inrae.fr/frenchcropusage/sparql> {
    ?body_fcu 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.

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.[33]

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-occuring 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[34] or its upper parent[35], we retrieve wheat

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

[34] http://purl.obolibrary.org/obo/NCBITaxon_318829

[35] http://purl.obolibrary.org/obo/NCBITaxon_639021

and rice genes co-occuring with these taxa. This may be the indication of orthologous genes in wheat and rice genomes that should be explored by domain experts.

```
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.

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[36] [MGTFZ17] to retrieve the alignments between NCBI classes and FCU concepts (line 20). Alignments between FCU concepts and TAXREF-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`[37] and `fcu:Bles_tendres_printemps`[38]. Note that this example query is meant to be executed on the WheatGenomicsSLKG SPARQL endpoint, invoking the two other SPARQL endpoints

[36]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

[37]http://ontology.inrae.fr/frenchcropusage/Bles_tendres_hiver

[38]http://ontology.inrae.fr/frenchcropusage/Bles_tendres_printemps

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.

```
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

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 independent from the NLP tools used for NERL and may be reused for representing annotations extracted using any NLP tools. The annotations rely on the same understanding of the OA model. Each annotation links one mention in a document to one domain-specific entity defined in external semantic resources. As illustrated in Figure 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).

Furthermore, 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 source resource. 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 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 extracted from the title, abstract or abstract's sub-parts, each having its own URI and being linked by the `frbr:partOf` property (Figure 7). 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.

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) [CIdD+20]. 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. Figure 10 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 Figure 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.

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

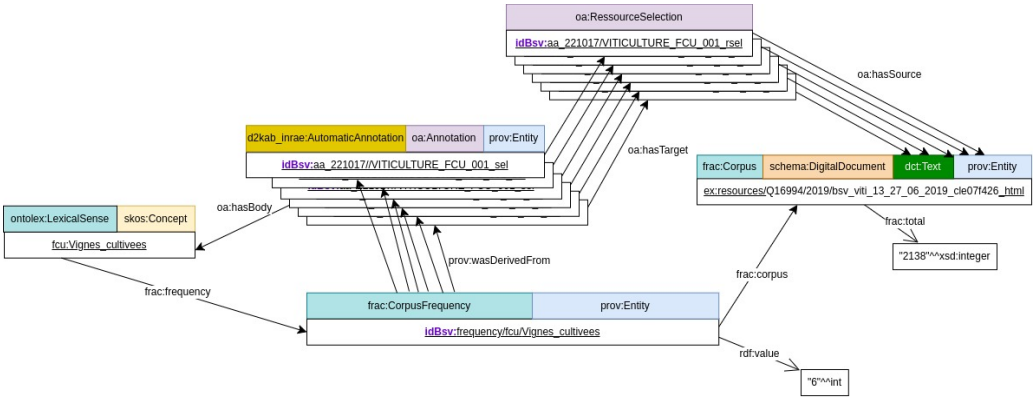


Figure 10: Frequency Modelling in PHB graph based on FRAC, OA and PROV vocabularies.

principles. We rely on Semantic Web models and technologies to build domain-specific KGs that allow domain experts 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 varities). 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 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.

On another note, we plan to construct and publish richly annotated gold standard datasets based on the three corpora. This will require considerable efforts of domain experts to define guidelines and samples for NE 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.

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 is 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 is 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 is 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.

References

[CIdD+20] Christian Chiarcos, Maxim Ionov, Jesse de Does, Katrien Depuydt, Fahad Khan, Sander Stolk, Thierry Declerck, and John Philip McCrae. Modelling frequency and attestations for ontolex-lemon. In Proceedings of the 2020 Globalex Workshop on Linked Lexicography, pages 1–9, 2020.

[DFMd19] Brett Drury, Robson Fernandes, Maria-Fernanda Moura, and Alneu de Andrade Lopes. A survey of semantic web technology for agriculture. Information Processing in Agriculture, 6(4):487–501, 2019.

[GF95] Michael Grüninger and Mark S. Fox. The Role of Competency Questions in Enterprise Engineering, pages 22–31. Springer US, Boston, MA, 1995.

[JTA+18] Clément Jonquet, Anne Toulet, Elizabeth Arnaud, Sophie Aubin, Esther Dzalé Yeumo, Vincent Emonet, John Graybeal, Marie-Angélique Laporte, Mark A. Musen, Valeria Pesce, and Pierre Larmande. Agroportal: A vocabulary and ontology repository for agronomy. Computers and Electronics in Agriculture, 144:126–143, 2018.

[KCD+22] 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, and Ciprian-Octavian Truica. When linguistics meets web technologies. recent advances in modelling linguistic linked data. Semantic Web, 13(6):987–1050, 2022.

[KY06] Nori Kurata and Yukiko Yamazaki. Oryzabase. An Integrated Biological and Genome Information Database for Rice. Plant Physiol., 140(1):12, January 2006.

[MDFM15] F. Michel, L. Djimenou, C. Faron-Zucker, and 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), pages 443–454, Lisbon, Portugal, 2015.

[Mei18] Uwe Meier. Growth stages of mono- and dicotyledonous plants: BBCH Monograph. Open Agrar Repositorium, 2018.

[MFZCG19] Franck Michel, Catherine Faron-Zucker, Olivier Corby, and 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, pages pp. 883–892, San Francisco, USA, 2019. ACM Press.

[MG13] Luc Moreau and Paul Groth. Provenance: an introduction to prov. Synthesis lectures on the semantic web: theory and technology, 3(4):1–129, 2013.

[MGA+20] 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, and 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, volume 12507 of Lecture Notes in Computer Science, pages 294–310. Springer, 2020.

[MGTFZ17] Franck Michel, Olivier Gargominy, Sandrine Tercerie, and 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), volume 1933, Vienna, Austria, 2017. CEUR Workshop Proceedings. http://ceur-ws.org/Vol-1933/paper-3.pdf.

[MRHLA20] Jose Martinez-Rodriguez, Aidan Hogan, and Ivan Lopez-Arevalo. Information extraction meets the semantic web: A survey. Semantic Web, vol. 11:pp. 255–335, 2020.

[NBV+14] Claire Nédellec, Robert Bossy, Dialekti Valsamou, Marion Ranoux, Wiktoria Golik, and Pierre Sourdille. Information extraction from bibliography for marker-assisted selection in wheat. In Proceedings of Research Conference on Metadata and Semantics Research - MTSR 2014, pages pp. 301–313, 2014.

[NIBS20] Claire Nédellec, Liliana L. Ibanescu, Robert Bossy, and Pierre Sourdille. WTO, an ontology for wheat traits and phenotypes in scientific publications. Genomics & Informatics, 18(2), 2020.

[Per16] Silvio Peroni. SAMOD: an agile methodology for the development of ontologies. In Mauro Dragoni, Maréa Poveda-Villalón, and Ernesto Jimenez-Ruiz, editors, OWL: Experiences and Directions – Reasoner Evaluation, pages 55–69. Springer, 2016.

[PG08] Valentina Presutti and Aldo Gangemi. Content ontology design patterns as practical building blocks for web ontologies. In International conference on conceptual modeling, pages 128–141. Springer, 2008.

[PS12] Silvio Peroni and David Shotton. FaBiO and CiTO: Ontologies for describing bibliographic resources and citations. Journal of Web Semantics, 17:33–43, 2012.

[RBP+17] Catherine Roussey, Stephan Bernard, François Pinet, Xavier Reboud, Vincent Cellier, Ivan Sivadon, Danièle Simonneau, and Anne-Laure Bourigault. A methodology for the publication of agricultural alert bulletins as lod. Computers and Electronics in Agriculture, 142:632–650, 2017.

[RDA+21] Catherine Roussey, Xavier Delpuech, Florence Amardeilh, Stephan Bernard, and Clement Jonquet. Semantic description of plant phenological development stages, starting with grapevine. In Emmanouel Garoufallou and María-Antonia Ovalle-Perandones, editors, Metadata and Semantic Research, pages 257–268, Cham, 2021. Springer International Publishing.

[SCD+20] 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, and Ilene Karsch-Mizrachi. Ncbi taxonomy: a comprehensive update on curation, resources and tools. Database : the journal of biological databases and curation, 2020, 2020.

[SCY17] Robert Sanderson, Paolo Ciccarese, and Benjamin Young. Web annotation ontology. Technical report, W3C, 2017.

[The22] The UniProt Consortium . UniProt: the Universal Protein Knowledgebase in 2023. Nucleic Acids Research, 51(D1):D523–D531, 11 2022.

[UG96] Mike Uschold and Michael Gruninger. Ontologies: Principles, methods and applications. The Knowledge Engineering Review, 11(2):93–136, 1996.

[VTNH+18] Aravind Venkatesan, Gildas Tagny Ngompe, Nordine El Hassouni, Imene Chentli, Valentin Guignon, Clement Jonquet, Manuel Ruiz, and Pierre Larmande. Agronomic Linked Data (AgroLD): A knowledge-based system to enable integrative biology in agronomy. PLOS ONE, 13(11):1–17, 2018.

[WDA+16] 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, and Barend Mons. The fair guiding principles for scientific data management and stewardship. Scientific Data, 3, 03 2016.

[WSM+20] Leon Weber, Mario Sänger, Jannes Münchmeyer, Maryam Habibi, Ulf Leser, and Alan Akbik. HunFlair: An Easy-to-Use Tool for State-of-the-Art Biomedical Named Entity Recognition. arXiv:2008.07347 [cs], August 2020. arXiv: 2008.07347.

[YBC+22] 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, and Taner Z Sen. Grain-

Genes: a data-rich repository for small grains genetics and genomics. Database, 2022, 05 2022. baac034.