



HAL
open science

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

► To cite this version:

Nadia Yacoubi Ayadi, Stephan Bernard, Robert Bossy, Marine Courtin, Bill Gates Happi Happi, et al.. A Unified Approach to Publish Semantic Annotations of Agricultural Documents as Knowledge Graphs. I3S, Université Côte d'Azur; Paris Saclay University; INRAE; IRD. 2024, pp.43. hal-04495022

HAL Id: hal-04495022

<https://hal.science/hal-04495022>

Submitted on 8 Mar 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



Distributed under a Creative Commons Attribution 4.0 International License

1 A Unified Approach to Publish Semantic Annotations of 2 Agricultural Documents as Knowledge Graphs

3 Nadia Yacoubi Ayadi^{1,2}, Stephan Bernard³, Robert Bossy⁴, Marine Courtin^{3,4},
4 Bill Gates Happi Happi⁵, Pierre Larmande⁵, Franck Michel¹, Claire Nedellec⁴,
5 Catherine Roussey^{3,6}, and Catherine Faron¹

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

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

8 ³TSCF, INRAE, Centre Auvergne Rhône Alpes Clermont, Aubière, France

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

10 ⁵DIADE, IRD, CIRAD, Univ. Montpellier, Montpellier, France

11 ⁶Mistea, INRAE, Centre Occitanie, Montpellier, France

12 January 31, 2024

13 **Abstract**

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

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

38 **Keywords**

39 Agriculture, Knowledge Graphs, Semantic modelling, RDF Transformation, Natural Lan-
40 guage Processing, Annotations, Semantic Resources, Named Entity Recognition and Link-
41 ing

42 **1 Introduction**

43 Knowledge Graphs (KG) are multi-relational graphs of relations between well-defined and
44 uniquely identifiable entities created from heterogeneous data sources. They enable to develop
45 data management platforms compliant with the FAIR (Findability, Accessibility, Interoperabil-
46 ity and Reuse) principles [WDA⁺16] referring to best practice guidelines: resources must be
47 accessible, understood, exchanged and reused by machines. A typical approach towards pub-
48 lishing FAIR knowledge graphs is to rely on Linked Data (LD) principles and Semantic Web

49 technologies (SWT). Indeed, RDF and other Semantic Web standards are designed to promote
50 interoperability and linking between datasets. Additionally, to ensure that RDF datasets are
51 truly interoperable and reusable within a specific field, they must rely on domain-specific and
52 open vocabularies, models, and data category registries capturing the shared theoretical foun-
53 dations and terminology used by a community of domain experts [KCD⁺22]. Constructing
54 knowledge graphs from unstructured data enables to bridge the gap between the huge amount
55 of heterogeneous data and easily explore and query it to address various use cases. This paper
56 focuses on building knowledge graphs from textual data sources by extracting relevant domain-
57 specific entities and organizing them in a structured and meaningful annotation. This approach
58 can be beneficial in making sense of large, complex and heterogeneous datasets, linking related
59 information and knowledge, and providing intuitive ways to access and explore domain data
60 and knowledge, adhering to the FAIR principles. We present a methodology for constructing
61 domain-specific knowledge graphs using SWT, which involves the re-use of shared RDF-based
62 vocabularies and models. Although the proposed methodology can be applied to various do-
63 mains and can support a wide range of use cases, in this paper we focus on building knowledge
64 graphs representing semantic annotations of textual documents in the field of agriculture. We
65 consider three different text corpora and we demonstrate how we leverage Natural Language
66 Processing (NLP) techniques to first extract different types of named entities and then struc-
67 ture and integrate them into KGs using the same data model. Two corpora are collections
68 of scientific publications on rice and wheat functional genomics, retrieved from the PubMed¹
69 repository. These publications investigate the gene-phenotype link for varietal selection, or
70 more precisely the identification of gene markers involved in the expression of a given pheno-
71 type, for selection assistance [NBV⁺14]. A third corpus gathers technical documents called
72 Plant Health Bulletins (PHBs) which are agricultural alert bulletins published in France. Al-
73 though, the KGs that we built from these corpora are intended to serve different needs, we
74 have adopted the same methodology and core data model. The first step of our methodology
75 leverages NLP pipelines to perform the tasks of Named Entity Recognition (NER) and Linking
76 (NEL) [MRHLA20]. A semantic data model has been defined to capture how NE annotations

¹<https://pubmed.ncbi.nlm.nih.gov/>

77 produced by NLP pipelines should be structured and described in each KG. We initiated the
78 data model definition with a set of competency questions, together with motivating examples.
79 We were inspired by the agile SAMOD methodology [Per16], which in turn is based on the early
80 work of Uschold & Gruninger [GF95, UG96]. The SAMOD process is initiated by a motivating
81 scenario which lead to a set of competency questions (CQs) that provide requirements on the
82 knowledge graphs to be created. All CQ demonstrate that the KGs should be uniformly queried
83 by experts to highlight the context of NE co-occurrence in the original texts in order to reveal
84 hidden interactions between NE. In the continuity of earlier works [MGA⁺20], we propose to
85 rely on the Open Annotation Ontology (OA) [SCY17] to describe, structure and integrate NE
86 annotations and their occurrence contexts in texts. Domain specific vocabularies are also reused
87 to describe bibliographic information of PubMed publications and provenance information for
88 the PHB corpus. The resulting model was automatically populated using a mapping-based
89 transformation pipeline implemented with the Morph-xR2RML tool [MDFM15].

90 The paper is structured as follows. in Section 2, we present the materials of our research
91 work which consist of three text corpora and the semantic resources used to annotate those
92 corpora. in Section 3.1, we present the CQs of each case study. We describe the proposed
93 semantic model in further details in Section 3.2. in Section 4, we present the validation results
94 of the case studies. Finally, in Section 5, we discuss the results and synthesize the learned
95 lessons before concluding in Section 6.

96 **2 Materials**

97 In this section, we present the different materials used in this research work to build our KGs. in
98 Section 2.1, we first present the text corpora that were processed using different NLP pipelines
99 in order to generate semantic annotations of domain named entities. For the linking task, NLP
100 processes rely on a set of semantic resources presented in Section 2.2.

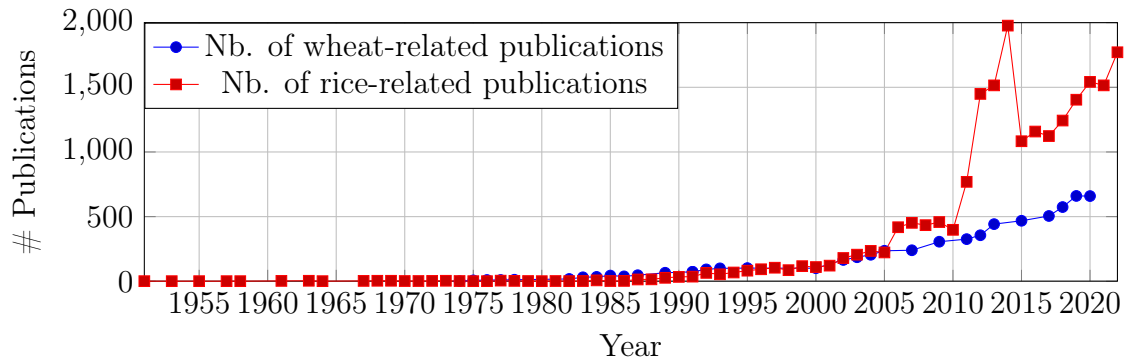


Figure 1: Evolution of the number of research works on wheat and rice genomics from 1951 to 2022

101 2.1 Text Corpora

102 2.1.1 PubMed Corpus on Wheat Genomics

103 The MaIAGE research group² has collected 8,496 scientific references of articles published
 104 between 1974 and 2021 related to wheat genomics wheat selection. A first exploration of the
 105 corpus shows that over the last two decades, the number of publications on wheat genomics
 106 has steadily increased. While the first publication appears in 1974, the annual publication grew
 107 slowly between 1975 and 2000. More than 80% of the publications in the corpus originate from
 108 the last two decades, which reflects the significant rise of research interest in wheat genomics
 109 as shown in Figure 1.

110 In this corpus, the PubMed identifier, title and abstract of each publication are provided.
 111 In several cases, the abstract of a publication is organised in different sub-sections. We used the
 112 AlvisNLP pipeline to identify genes, traits, phenotypes, taxa and varieties entities mentioned
 113 in the title and the abstract of publications, and the relationships between wheat varieties and
 114 phenotypes. In total, 88,880 mentions of 4,318 distinct named entities were recognized and
 115 linked to exiting entities of semantic resources (presented in Section 2.2). Figure 4 illustrates
 116 an example of PubMed publication where three types of NEs are recognised: we distinguish
 117 between NE mentions that refer to genes (e.g., *Sr2*, *Lr27*, *Lr34*), phenotypes (e.g., *leaf rust*
 118 *resistance*, *resistance to stem rust*, *powdery mildew resistance*) and taxa (e.g., *wheat*).

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

²<https://maiage.inrae.fr/en>

121 2.1.2 PubMed Corpus on Rice Genomics

122 The DIADE research group³ collected 17,058 scientific articles from the Oryzabase database [KY06]
123 which provides manually checked PubMed entries related to rice genomics. The corpus rep-
124 resents scientific articles published between 1951 and 2021. It is worth noticing an increasing
125 pace of publishing activity during this period, first coinciding with the availability of the two
126 main rice genomes and their annotations (2004-2008) and then to the development of second
127 and third generation of sequencing techniques (2012-2018) allowing faster and cheaper genome
128 sequences availability. Most of the articles in the corpus date from the past two decades, which
129 is consistent with the sharp increase in research interest in rice genetics as depicted in Figure 1.
130 We used the HunFLAIR NER tagger [WSM⁺20] that we combined with NLTK, Spacy and
131 other Python libraries to extract four types of named entities in the title or the abstract of the
132 articles: we distinguish between NE mentions that refer to genes (e.g., *OsMAPK2* or *MOC1*),
133 species (e.g., *Oryza sativa* or *Magnaporthe oryzae*), chemicals (e.g., *gibberellic acid* or *nitrogen*)
134 and diseases or phenotypes (e.g., diseases *blast* or *Sheath blight disease*). In total, 351,003 men-
135 tions of 63,591 distinct NEs were identified from PubMed abstracts and titles. When possible,
136 these NEs were linked with existing semantic resources as explained in Section 2.2.

137 2.1.3 Plant Health Bulletin Corpus

138 In France, the Grenelle Environment and Ecophyto 2018 program strengthened national surveil-
139 lance networks of crops and agricultural practices. Plant Health Bulletins are one of the modal-
140 ities established by these surveillance networks in all regions and French overseas departments.
141 A Plant Health Bulletin (PHB) is an agricultural alert document, both technical and regulatory
142 in nature, written in French under the responsibility of a regional epidemiological surveillance
143 committee. A PHB gathers information about the health status of crops. It reports observa-
144 tions of crop development and pest attacks, and analyses pest risk in the whole area.

145 Nearly 15,000 plots are observed each year to edit approximately 3400 PHBs per year
146 [RBP⁺17]. PHBs synthesize the interpretation of observations performed on crops by different
147 collecting networks, elements from epidemiological models, meteorological data and sometimes

³<http://diade.ird.fr/>

148 biological analysis. Thus, the PHB corpus can be seen as a French archive of human validated
149 crop observations on the whole French territory. The TSCF research group has collected 36,469
150 bulletins from 2009 to 2022 from the whole French territory. In this work, we considered three
151 sub-corpora of PHBs previously used to validate NLP processes: the Vespa corpus gathers 500
152 PHBs collected in the whole French territory between 2009 and 2015; the D2KAB corpus is
153 composed of 230 PHBs collected in 2019, manually selected to cover the whole French territory
154 and to represent three crop categories – field crops, vegetables and grapevines; and the Alea
155 corpus is composed of 150 PHBs randomly selected from the whole corpus, that is to say,
156 the publication date may vary from 2009 to 2020. These three sub-corpora are available on
157 a Git repository⁴. The whole corpus brings together a total of 880 PHBs with an average of
158 2,548 tokens per bulletin, covering the whole French territory and all crop categories of French
159 agriculture. We use the AlvisNLP pipeline to extract NEs referring to french crop names and
160 french development stages mentioned in the text of PHBs. In total, 37,488 mentions of 416
161 distinct NEs were extracted. Figure 2 illustrates an example of PHB where two types of NEs
162 are recognised. We distinguish NE mentions that refer to french crop names: e.g., *viticulture*,
163 *fleurs* (flowers), *baies* (berries), *Melon*, *pois* (peas); and french development stages : e.g.,
164 *floraison* (flowering), *BBCH 69*, *BBCH-69 et 73* , *BBCH-75*, *développement des fruits* (fruit
165 development). Those mentions are linked to existing elements defined in the FCU thesaurus
166 and the BBCH-based Plant Phenological Description Ontology.

167 2.2 Semantic Resources

168 In the agriculture domain, an increasing number of semantic resources (ontologies, thesaurii)
169 was developed and published using Semantic Web technologies [DFMd19] and made available
170 for research communities in open portals such the Agroportal repository⁵ [JTA⁺18]. In this
171 section, we present the semantic resources that we have reused to annotate the text corpora
172 presented in Section 2.1.

⁴<https://forgemia.inra.fr/bsv/corpus-bsv>

⁵<http://agroportal.lirmm.fr/>

ACTUALITES
<p>Phénologie Fin floraison à nouaison.</p>
<p>Vers de la grappe Glomérules vides, vol de 2ème Génération imminent.</p>
<p>Mildiou Sorties de symptômes sur les témoins, situation toujours saine.</p>
<p>Oïdium Doucement mais sûrement...vigilance à maintenir.</p>
<p>Cicadelles vertes Populations encore non pré-</p>

Phénologie

• Nouaison en cours.

La floraison s'est nettement accélérée depuis le week-end dernier sur le vignoble. Les stades oscillent entre fin floraison (BBCH 69 80% de fleurs ouvertes) et 71 (nouaison) sur l'Aubance et le Layon. Le Saumurois, le Sèvre et Maine et le Pays de Retz se situaient en début de semaine entre les stades BBCH-69 et 73 (grains de plomb, baies 2-3 mm).

Les parcelles les plus précoces (Chardonnay, Gamay, Pinot gris, Melon B) du réseau sont presque au stade BBCH-75 (petit pois) mais souvent de façon hétérogène.

L'hétérogénéité paraît cependant un peu s'estomper actuellement avec la pousse en accéléré du week-end



Stade 71—nouaison : début de développement des fruits, les déchets floraux sont tombés.

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

173 2.2.1 Wheat Trait and Phenotype Ontology

174 The Wheat Trait and Phenotype Ontology (WTO) [NIBS20] is a domain ontology that covers
175 a wide range of wheat traits and phenotypes related to soft wheat (*Triticum aestivum L.*) and
176 the environmental factors that affect these traits. While traits denote physical observable plant
177 properties, phenotypes are the set of possible values of traits. Capturing phenotypic information
178 in a formal, shared representation is crucial for scientists as well as for breeders. However,
179 automatic annotation of textual data remains a challenge due to the large number of traits and
180 the great diversity of the vocabulary used to designate them. WTO has been developed to meet
181 the requirements of trait and phenotype annotation in the scientific literature. In WTO, traits
182 are organised into different categories such as development, morphology, quality, response to
183 environmental conditions including biotic and abiotic stresses.

184 WTO (3.0) is available in the OBO format on Agroportal⁶ and contains 745 classes. We
185 revised and transformed WTO 3.0 in OWL/SKOS format⁷ and we used it to annotate mentions
186 of phenotypes and traits recognised in the PubMed corpus on wheat functional genomics.

187 2.2.2 NCBI Taxonomy

188 The NCBITaxon ontology⁸ is an automatic translation of the NCBI taxonomy into OWL.
189 The NCBI Taxonomy consists of a single, hierarchically arranged list of organismal names
190 across all domains of life. These names are correct, current and valid according to the best
191 authorities within the separate taxonomic disciplines and codes of nomenclature [SCD⁺20]. In
192 the NCBITaxon ontology, the NCBI taxons are translated into OWL classes whose instances
193 would be individual organisms. The labels of NCBITaxon classes are the scientific names
194 (e.g. *Triticum aestivum L.*) and vernacular names (e.g. *soft wheat*) of the taxons. We used
195 NCBITaxon to annotate and link different types of organisms mentioned in both PubMed
196 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>

197 **2.2.3 French Crop Usage Thesaurus**

198 The French Crop Usage (FCU) thesaurus⁹ organises plants based on their roles in agriculture,
 199 or in other words, agricultural plant uses. The thesaurus hierarchy has two main branches as
 200 shown in Figure 3. The branch named *Multiusages* contains all the cultivated plants that have
 201 several uses in agriculture. For example, *carotte* (carrot) may be used as vegetable or as fodder.
 202 The branch *Usages_plantes_cultivees* organises cultivated plants according to their uses and
 203 represents crop categories. FCU stores only the french vernacular names of plants. The FCU
 204 thesaurus is formalized using SKOS and used in this work to extract and link crop names in
 205 the PHB corpus.

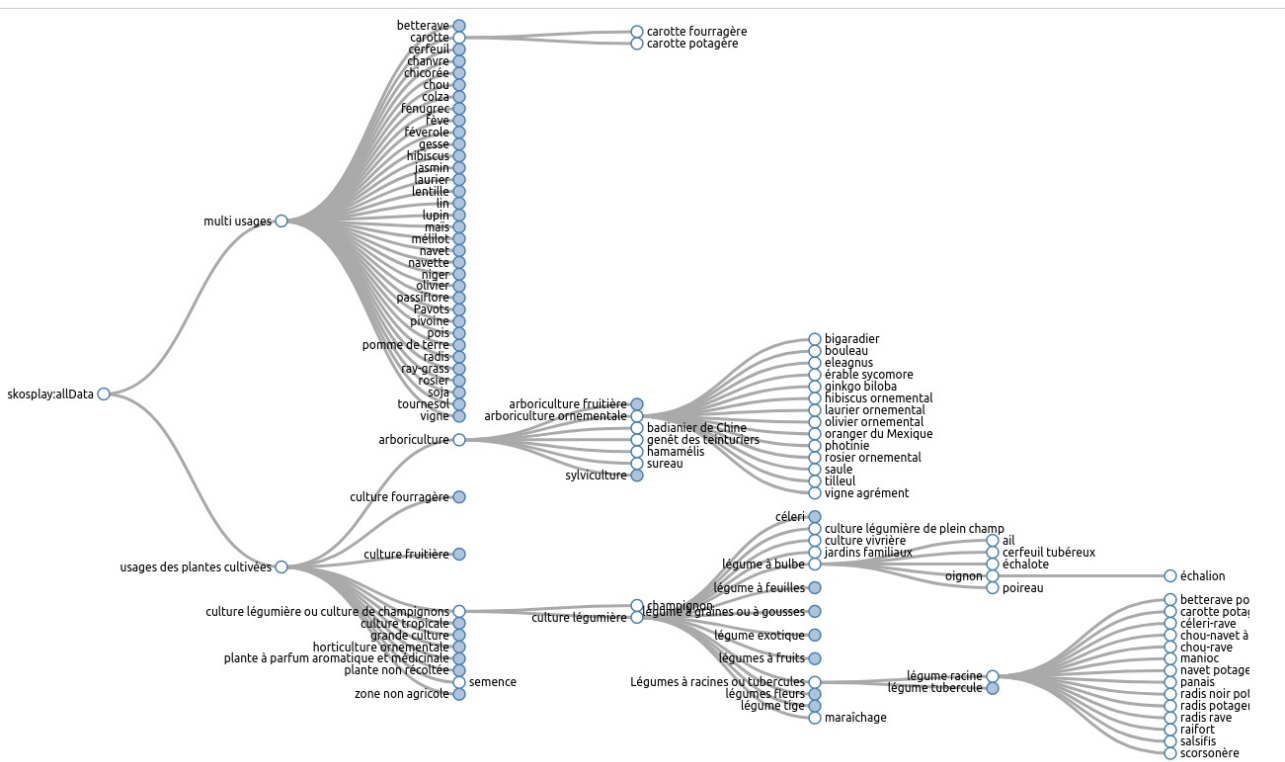


Figure 3: An extract from the FCU thesaurus. Visualisation generated by the SKOS Play tool.

206 **2.2.4 BBCH-based Plant Phenological Description Ontology**

207 BBCH (*Biologische Bundesanstalt, Bundessortenamt und CHemische Industrie*) is considered
 208 as a reference to describe development stages of different plant species in four languages: En-
 209 glish, French, Spanish and German. It describes several sets of development stages. A set of

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

210 stages composes a BBCH scale. Some plant species (like tomatoes or potatoes) have a specific
211 set of stages named 'individual scales'. A general BBCH scale is also defined for plant species
212 where no individual scale exists [Mei18]. BBCH framework uniformly codes phenologically
213 similar development stages of different plant species. The BBCH-based Plant Phenological De-
214 scription Ontology (PPDO)¹⁰ [RDA⁺21] relies on the BBCH scale. It formalizes the scales and
215 their associated stages as a specialization of the SKOS model. A stage is an instance of class
216 *skos:Concept* with labels and definitions in four languages. The BBCH scale is represented as
217 a SKOS thesaurus.

218 2.2.5 Resources for other mentions

219 Resources for genes, markers, wheat and rice varieties published as LOD datasets are very
220 limited and in most cases, they are either incomplete, do not come from authoritative orga-
221 nizations, or do not provide unique identifiers. Among the available semantic resources for
222 genes and markers, the UniProt Knowledge base [The22] (UniProtKB) is a central hub for a
223 collection of functional information on proteins, with accurate, consistent and rich annotation
224 which is accessible through a SPARQL endpoint¹¹. However, multiple UniProtKB identifiers
225 can be retrieved for the same genomic entity which makes it impossible to link named enti-
226 ties using this resource. Therefore, in order to recognize and normalize genes and markers
227 from texts, AlvisNLP and HunFLAIR both rely on curated domain lexicons or dictionaries
228 combined with patterns. For wheat genes and markers, a curated list of gene names from the
229 GrainGenes [YBC⁺22] database was created. For rice genes, the Oryzabase [KY06] database
230 was used and integrated into the AgroLD Knowledge Graph [VTNH⁺18] which capitalises ge-
231 nomic data about plant species of high interest for the plant science community (among which
232 rice and wheat) to provide functional information on genes and their relationship across species.
233 AgroLD is available through a SPARQL endpoint¹².

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

¹⁰The 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>

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

238 **3 Method**

239 **3.1 Competency Questions**

240 In this section, we present a set of Competency Questions (CQs) stemming from requirements
241 expressed by experts and collected in the context of the D2KAB project¹⁵. CQs are natural
242 language questions illustrating the typical knowledge that scientists would require a data source
243 to provide. A common way of validating a KG is to provide the formalisation of CQs, for a
244 given case study, as SPARQL queries using the KG model. In the following we present the CQs
245 for our case studies. Their formalisation in SPARQL is presented in Section 4.

246 **3.1.1 Competency Questions for Scientific Literature Exploration in Wheat and** 247 **Rice functional genomics**

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

257 Hence, exploiting the ever-growing scientific literature (Figure 1) could help scientists to
258 discover hidden interactions between entities of interest for functional genomics by examining
259 their co-occurrence in scientific publications. Thus, structuring and integrating genomic NEs
260 extracted from scientific publications and annotated based on relevant knowledge from external

¹⁴<https://www.geves.fr/catalogue/>

¹⁵<https://www.d2kab.org/>

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 Spielmeier

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

261 semantic resources is essential. Considering the PubMed publications corpus presented in
262 Section 2, we present here a subset of CQ that ultimately consists of a set of research questions.

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

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

276 *CQ2: Which genetic markers appear proximal to a specific gene, and which genes are men-*
277 *tioned proximal to a particular phenotype in publications dating from after 2010?*

278 The CQ2 is designed for the PubMed corpus on wheat genomics, since the NEs of the
279 genetic markers are recognized only in this corpus. A genetic marker discriminates the different
280 alleles of a gene with the polymorphism of the DNA sequence. Thus, genetic markers are used

281 to select the wheat varieties with a trait or phenotype of agronomic interest [NBV⁺14]. For
282 instance, in the case of *resistance to the stripe rust disease* in wheat, the gene *Yr65* is often
283 mentioned in literature along with this phenotype. Furthermore, markers such as *Xgdm33*,
284 *Xgwm11*, *Xgwm18*, and *Xgwm413* are mentioned in the same context as this gene. As the
285 techniques for genetic markers selection have evolved over time and some of them have become
286 obsolete, the expert can also refine the query to select only publications which appeared after
287 2010. The knowledge graph should contain the publication metadata such as publication year,
288 list of authors, or the number of incoming citations.

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

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

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

298 CQ3 reflects the need for experts to conduct a systematic literature review of publications
299 that mention specific genes cited in the literature proximal to certain traits (from a specific
300 family of traits) as well as wheat or rice varieties. The results of this query should include a list
301 of articles mentioning, in their abstracts or titles, gene names, a wheat or rice variety and a set
302 of traits known, for instance, to be involved in pathogen resistance. For instance, a scientist
303 may be interested in resistance to fungal pathogens which cause massive and destructive losses
304 to crops. Thus, the study of resistance mechanisms is essential to fully understand the inter-
305 actions between pathogens across crop varieties. Based on the WTO structure which classifies
306 traits in different taxonomies, it is possible to conduct this study for all traits belonging to
307 the sub-hierarchy of fungal pathogen resistance class. This CQ highlights the importance to
308 incorporate domain knowledge formally represented in ontological and terminological resources
309 (e.g., WTO).

310 *CQ4: Which gene names are cited in the literature proximal to a specific taxon (and option-*
311 *ally to one or more of its descendants)?*

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

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

317 It has been demonstrated that some fungal and bacterial disease pathogens affect both rice
318 and wheat. Wheat and rice disease resistance has been studied for a large panel of pathogens,
319 including *rusts*, *smuts*, *Fusarium head blight*, *Septoria leaf blotch*, *tan spot*, and *powdery mildew*,
320 that cause the most serious losses. The goal is to search for wheat and rice genes co-occurring
321 in literature with the same taxon of a pathogen (or a more specific taxon). This enables to
322 identify orthologous genes¹⁶ in wheat and rice.

323 **3.1.2 Competency Questions for Agronomic Studies**

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

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

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

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

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

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

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

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

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

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

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

358 *CQ9: What is the scientific literature available on the crop to which a PHB bulletin relates?*

359 The goal is to identify which new research publications is related to a crop cultivated in the
360 French territory, to search for example new crop varieties that are resistant to drought or high
361 temperatures.

362 **3.2 Proposed Semantic Model**

363 We reuse a set of state-of-the-art vocabularies to design a unified semantic model that captures
364 the context of occurrence of several types of NE annotations in documents. The core part of

365 this model leverages and extends the model previously proposed in [MGA⁺20]. It is based on
366 the W3C Web Annotation Ontology (OA) [SCY17] to structure, describe and integrate NEs
367 extracted from both corpora, and eight complementary vocabularies to describe documents and
368 NEs. Table 1 shows the main vocabularies used to describe named entities annotations as well
369 as documents in both corpora.

370 3.2.1 OA-Based Model for Text Annotations with Named Entities

371 The Web Annotation Data Model is an ontology for structuring and sharing any type of an-
372 notations in an interoperable format. According to the OA documentation¹⁷, “*an annotation*
373 *is considered to be a set of connected resources (each identified by a URI), typically comprising*
374 *a body and a target where the body is somehow about the target*”. The core OA data model is
375 that an annotation a_i is an instance of the `oa:Annotation` class such that:

- 376 • The `oa:hasTarget` property identifies the part of document that is being annotated
377 with annotation a_i . The target is a resource selection with a selector, i.e., a resource
378 that identifies the part of text m_e that mentions a recognized entity e . In this work,
379 we use different types of selectors: `oa:TextQuoteSelector`, `oa:TextPositionSelector`,
380 `oa:XPathSelector` to indicate respectively the NE’s mention m_e (i.e., surface form), the
381 start and end offset position of m_e in the text and/or the XPath expression to retrieve
382 m_e in the HTML structure of the text. The `oa:hasSource` property is used to specify
383 the URI of the source where the selector is applied, the source being either the URI of
384 the document or one of its sub-parts.
- 385 • The `oa:hasBody` property identifies the entity e defined in a domain vocabulary such as
386 WTO, NCBI taxonomy, PPDO or FCU thesaurus.

387 Figure 5 illustrates an example RDF graph that captures five instances of NE annotations
388 recognised in the title and the abstract of a publication in the PubMed corpus¹⁸. The title and
389 the abstract of the publication are identified by a URI and become the source of the target
390 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/>

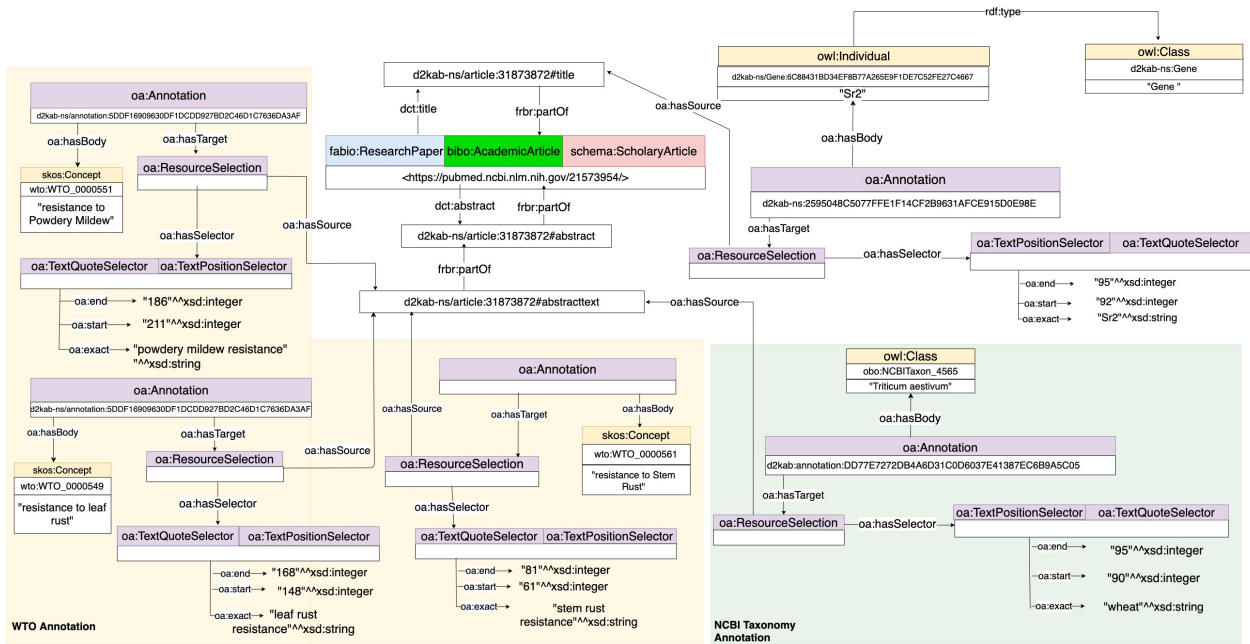


Figure 5: Example of NE annotations identified in a PubMed Publication's section (title and abstract) and represented in WheatGenomicsSLKG based on OA ontology.

391 in Figure 5). One annotation has as body a class from the NCBI taxonomy. One annotation
 392 has as body a gene entity URI that we have created locally in our graph (green area in Figure
 393 5). All mentions are identified by two selectors:

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

397 Figure 6 represents three annotations extracted from the PHB¹⁹ presented in Figure 2. One
 398 annotation²⁰ identifies the mention *Viticulture* localized in the main title of the PHB. Three
 399 types of selectors are used:

- 400 • an instance of `oa:XpathSelector` is used to express that the mention is found in the first
 401 section of the HTML element of type H1 which is a first level title.
- 402 • an instance of `oa:TextQuoteSelector` is used to specify the text of the mention, its
 403 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

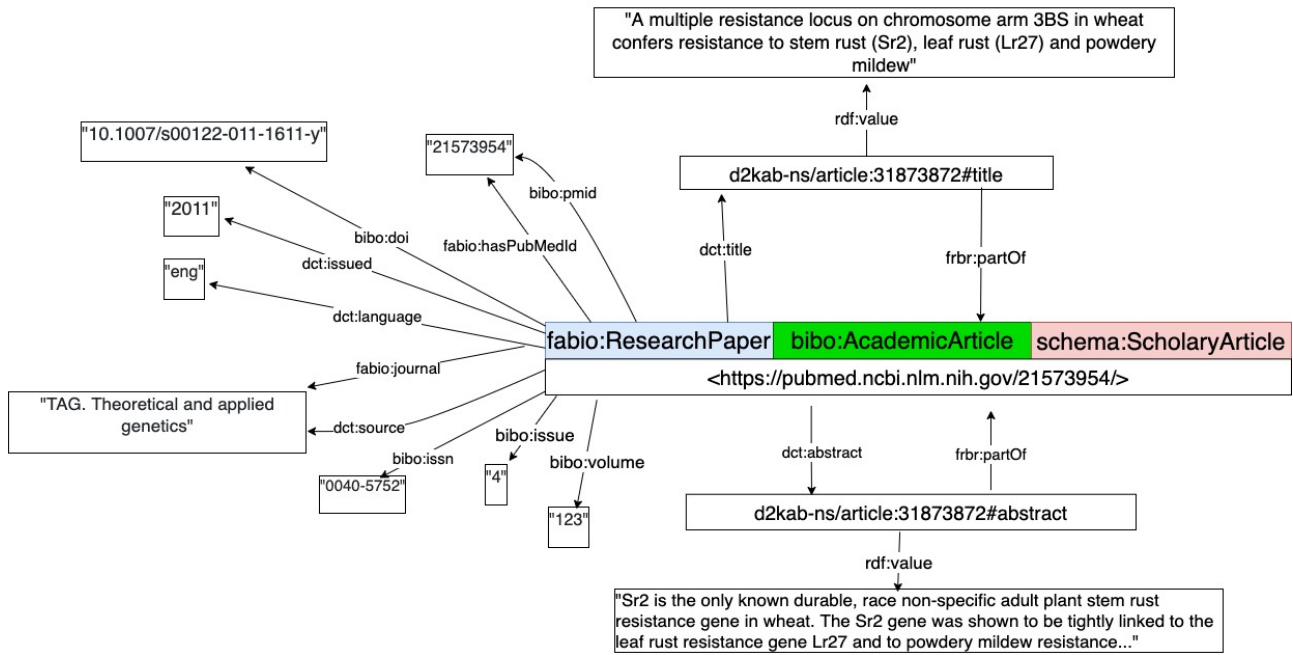


Figure 7: Example RDF graph describing bibliographic metadata of a scientific publication in the PubMed corpus

418 [MG13], the module for FRequency, Attestation and Corpus information (FRAC) of the LEXi-
 419 con Model for ONtologies (LEMON) [CiDD⁺20] and Schema.org. These vocabularies have been
 420 used slightly differently for each corpus.

421 **Bibliographic Metadata of PubMed Scientific Documents** For the PubMed corpus,
 422 we have reused bibliographic metadata vocabularies to describe specific attributes of scientific
 423 documents such as DOI, year of publication, number of pages, journal, etc. First, a scientific
 424 article is represented as an instance of classes `fabio:ResearchPaper`, `bibo:AcademicArticle`
 425 and `schema:ScholarlyArticle`. The Dublin Core properties `dct:title` and `dct:abstract`
 426 link the document to its title and abstract. Note that, in the PubMed corpus, abstracts may
 427 be structured in three subsections distinguished in our model by three different resources, each
 428 one identified by a unique URI. Property `frbr:partOf` is used to link an abstract and the
 429 document it is related to, or an abstract and one of its sub-sections. Figure 7 illustrates (a
 430 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

431 **Bibliographic Metadata for PHB Technical Documents** In the PHB corpus, a bulletin
432 has two digital realizations: a PDF file and a HTML file. Therefore, to model the biblio-
433 graphic information, we have reused an ontology design pattern from the DUL ontology called
434 `dul:InformationObject` [PG08]. An information object represents the generic information
435 about a document such as its publication date, the corpus to which it belongs, its associated
436 French region, its description. An information object has several realizations represented by
437 using property `dul:isRealizedBy`.

438 Figure 8 presents the graph annotating the bulletin of Figure 2.

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

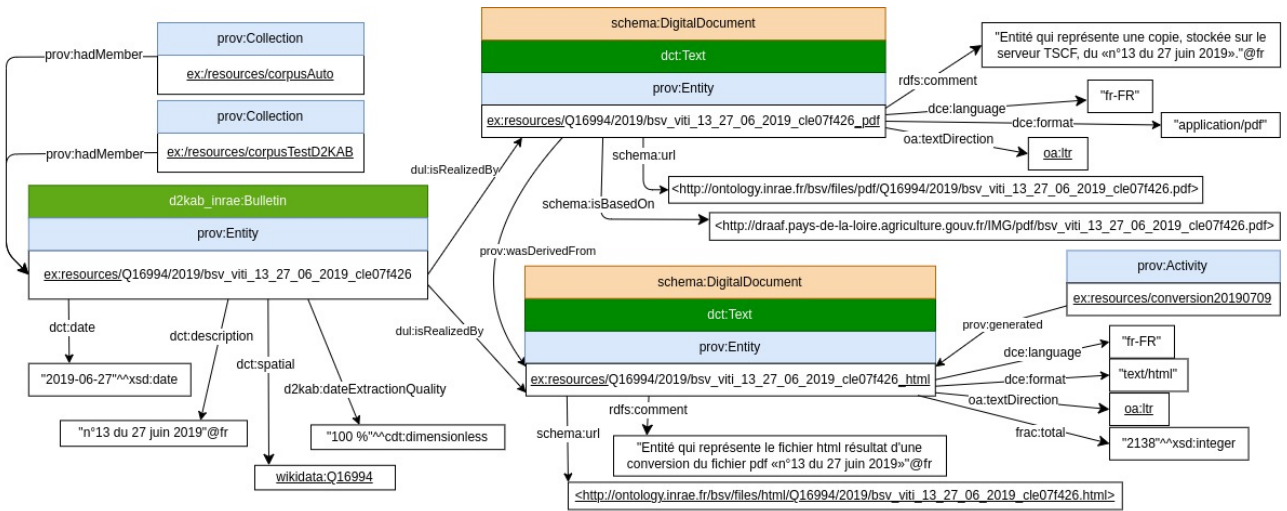


Figure 8: Example RDF graph describing metadata of a bulletin in the PHB corpus.

450 tokens.

451 3.2.3 Provenance Metadata

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

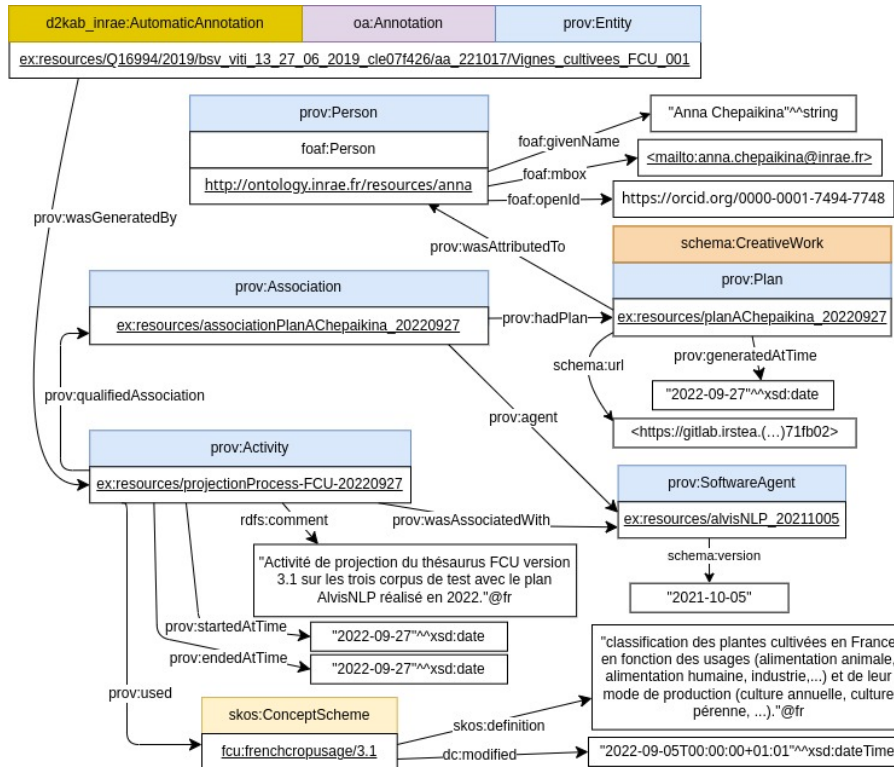


Figure 9: Example RDF graph describing provenance metadata of a bulletin in the PHB corpus.

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-xR2RML²³. 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	<code>http://ns.inria.fr/d2kab/{EntityClass}/{EntityID}</code>
Article	<code>https://pubmed.ncbi.nlm.nih.gov/{PubmedId}</code>
Annotation	<code>http://ns.inria.fr/d2kab/annotation/{annotationId}</code>
Title	<code>http://ns.inria.fr/d2kab/article/{PubmedId}#title</code>
Abstract	<code>http://ns.inria.fr/d2kab/article/{PubmedId}#abstract</code>
Abstract section	<code>http://ns.inria.fr/d2kab/article/{PubmedId}#{sectionName}</code>
Relation	<code>http://ns.inria.fr/d2kab/relation/{relationId}</code>

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

476 directory²⁴. Similar mapping rules have been defined to materialize the knowledge graph de-
 477 scribing the scientific literature on rice genomics, RiceGenomicsSLKG; they are available in
 478 the project’s GitHub directory²⁵. Table 2 illustrates the templates used to generate significant
 479 URIs for different types of resources in WheatGenomicsSLKG and RiceGenomicsSLKG.

480 In addition, in order to enrich the scientific publications with bibliographic metadata, we
 481 developed a SPARQL micro-service [MFZCG19] to query the PubMed Central API and retrieve
 482 publication metadata²⁶. For each publication, the micro-service transforms PubMed API’s
 483 results into an RDF graph that we insert in the KG being constructed. Finally, we also inserted
 484 as a subgraph of WheatGenomicsSLKG the SKOS version of the WTO semantic resources
 485 used to annotate phenotypes entities. WheatGenomicsSLKG and RiceGenomicsSLKG can be
 486 queried at <http://d2kab.i3s.unice.fr/sparql>. A list of SPARQL queries which implement CQs
 487 (CQ1 to CQ5) are presented in Section 4.2.

488 4.1.2 KG Pipeline for Plant Health Bulletins

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

493 A web-crawler is periodically run over the DRAAF websites to look for new PHBs that
 494 are downloaded while some information are extracted (download date, download URL, local
 495 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>

²⁶https://sparql-micro-services.org/service/pubmed/getArticleByPMId_sd/

496 The downloaded pdf files are transformed into HTML using the `pdf2blocks`²⁷ conversion tool.
497 AlvisNLP pipelines are used to extract NEs from these HTML files. Finally, the CSV output
498 files are transformed using specific xR2RML mapping rules. All the elements of this workflow
499 are available in the project gitlab repository²⁸. The resulting RDF KG can be queried at
500 <http://ontology.inrae.fr/bsv/sparql>. A list of SPARQL queries which implement CQs (CQ6 to
501 CQ10) are presented in Section 4.2.

502 4.2 Implementation of Competency Questions

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

508 4.2.1 SPARQL Queries Implementing CQs on the Wheat and Rice PubMed Cor- 509 pora

510 The SPARQL queries implementing CQs on PubMed corpora of wheat and rice functional
511 genomics can be executed at the SPARQL endpoint²⁹. The queries and excerpt of the obtained
512 results are provided as part of the supplementary materials. A Jupyter Notebook of these
513 SPARQL queries is available on our github repository³⁰.

514 **CQ1** : The SPARQL query presented in Listing 1 implements CQ1 and allows scientists to
515 retrieve genes that are mentioned proximal to the *resistance to leaf rust* trait considering the
516 WheatGenomicsSLKG graph. The query returns all genes mentioned proximal to the WTO
517 concept (`wto:0000483`) that corresponds to the aforementioned trait and counts the number
518 of times that a gene and the trait are recognized in the same context. The results of this query
519 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>

```

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.

520 also as the most frequent genes.

521 **CQ2** and **CQ2-bis** : The SPARQL query presented in Listing 2 implements CQ2 and
522 allows to identify genetic markers and genes mentioned proximal to a specific wheat trait in
523 scientific publications. The results of this query returns a list of scientific publications from
524 the WheatGenomicsSLKG graph that list several genetic markers and genes entities mentioned
525 proximal to the *resistance to Stripe Rust* trait. On another side, the SPARQL query presented
526 in Listing 3 corresponds to the implementation of CQ2-bis and allows scientists to retrieve gene
527 names that are mentioned proximal to the *GDP* chemical component in the scientific literature
528 on rice genomics.

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

534 **CQ4** : A first implementation of this CQ is presented in Listing 5 that performs a search
535 of gene mentions cited proximal to a specific taxon identified by a class in the NCBITaxon

```

1 SELECT (GROUP_CONCAT(distinct ?GeneName; SEPARATOR="-") as ?genes)
2 (GROUP_CONCAT(distinct ?marker; SEPARATOR="-") as ?markers) ?paper ?year ?WTOtrait
3 FROM NAMED <http://ns.inria.fr/d2kab/graph/wheatgenomicsslkg>
4 FROM NAMED <http://ns.inria.fr/d2kab/ontology/wto/v3>
5 WHERE {
6 VALUES ?WTOtrait { "resistance to Stripe rust" }
7 GRAPH <http://ns.inria.fr/d2kab/graph/wheatgenomicsslkg> {
8 ?a1 a oa:Annotation ;
9 oa:hasTarget [ oa:hasSource ?source1 ] ;
10 oa:hasBody [ a d2kab:Gene ; skos:prefLabel ?GeneName ] .
11 ?source1 frbr:partOf+ ?paper .
12 ?a2 a oa:Annotation ;
13 oa:hasTarget [ oa:hasSource ?source2 ] ;
14 oa:hasBody [ a d2kab:Marker ; skos:prefLabel ?marker ] .
15 ?source2 frbr:partOf+ ?paper .
16 ?a3 a oa:Annotation ;
17 oa:hasTarget [ oa:hasSource ?source3 ] ;
18 oa:hasBody ?WTOtraitURI .
19 ?source3 frbr:partOf+ ?paper .
20 ?paper a fabio:ResearchPaper ; dct:title ?source3 ; dct:issued ?year .
21 FILTER (?year >= "2010"^^xsd:gYear) }
22 GRAPH <http://ns.inria.fr/d2kab/ontology/wto/v3> {
23 ?WTOtraitURI skos:prefLabel ?WTOtrait . }
24 }
25 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*

```

1 SELECT ?GeneName (count(distinct ?paper) as ?NbOcc)
2 WHERE {
3 ?a1 a oa:Annotation;
4 oa:hasTarget [ oa:hasSource ?source1 ] ;
5 oa:hasBody [ a d2kab:Chemical; skos:prefLabel "GDP" ] .
6 ?source1 frbr:partOf+ ?paper .
7 ?a a oa:Annotation ;
8 oa:hasTarget [ oa:hasSource ?source ] ;
9 oa:hasBody [ a d2kab:Gene; skos:prefLabel ?GeneName ] .
10 ?source frbr:partOf+ ?paper .
11 ?paper a fabio:ResearchPaper .
12 }
13 GROUP BY ?GeneName
14 HAVING (count(distinct ?paper) > 0)
15 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

```

1 SELECT distinct ?paper ?Title ?GeneName ?varietyName ?WT0trait
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 [ a d2kab:Gene; skos:prefLabel ?GeneName ] .
9         ?source1 frbr:partOf+ ?paper .
10        ?a2 a oa:Annotation ;
11            oa:hasTarget [ oa:hasSource ?source2 ] ;
12            oa:hasBody ?body .
13        ?source2 frbr:partOf+ ?paper .
14        ?a3 a oa:Annotation ;
15            oa:hasTarget [ oa:hasSource ?source3 ] ;
16            oa:hasBody [ a d2kab:Variety; skos:prefLabel ?varietyName ] .
17        ?source3 frbr:partOf+ ?paper .
18        ?paper a fabio:ResearchPaper ; dct:title ?titleURI .
19        ?titleURI rdf:value ?Title .
20    GRAPH <http://ns.inria.fr/d2kab/ontology/wto/v3> {
21        { ?body skos:prefLabel ?WT0trait ; a ?class .
22          ?class rdfs:subClassOf* <http://opendata.inrae.fr/wto/0000340> . }
23    UNION
24    { ?body rdfs:label ?WT0trait ;
25      rdfs:subClassOf* <http://opendata.inrae.fr/wto/0000340> . }
26    UNION
27    { ?body skos:prefLabel ?WT0trait ; skos:broader* ?concept .
28      ?concept a ?class .
29      ?class rdfs:subClassOf* <http://opendata.inrae.fr/wto/0000340> . } }
30 } 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.

```

1 SELECT distinct ?paper ?title (GROUP_CONCAT(distinct ?geneName; SEPARATOR="-") as ?genes)
2     ?ncbiTaxon
3 FROM NAMED <http://ns.inria.fr/d2kab/graph/wheatgenomicsslkg>
4 FROM NAMED <http://purl.obolibrary.org/obo/ncbitaxon/ncbitaxon.owl>
5 WHERE {
6     VALUES ?ncbiTaxonURI {<http://purl.obolibrary.org/obo/NCBITaxon_208348>}
7     GRAPH <http://ns.inria.fr/d2kab/graph/wheatgenomicsslkg> {
8         ?a1 a oa:Annotation ;
9             oa:hasTarget [ oa:hasSource ?source1 ] ;
10            oa:hasBody [ a d2kab:Gene; skos:prefLabel ?geneName ] .
11        ?source1 frbr:partOf+ ?paper .
12        ?a3 a oa:Annotation ;
13            oa:hasTarget [ oa:hasSource ?source2 ] ;
14            oa:hasBody ?ncbiTaxonURI .
15        ?source2 frbr:partOf+ ?paper .
16        ?paper a fabio:ResearchPaper; dct:title ?titleURI .
17        ?titleURI rdf:value ?title . }
18    GRAPH <http://purl.obolibrary.org/obo/ncbitaxon/ncbitaxon.owl> {
19        ?ncbiTaxonURI rdfs:label ?ncbiTaxon . }
20 } 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.

536 ontology. Different taxa mentions can be also identified proximal to genes mentions in scientific
537 publications in both wheat and rice corpora. The SPARQL query presented in Listing 6 extends
538 the search for all sub-classes of a specific NCBITaxon class (*Puccinia*³¹).

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

```

1 SELECT distinct ?paper ?title (GROUP_CONCAT(distinct ?geneName; SEPARATOR="-") as ?genes) ?ncbiTaxon
2 FROM NAMED <http://ns.inria.fr/d2kab/graph/wheatgenomicsslkg>
3 FROM NAMED <http://purl.obolibrary.org/obo/ncbitaxon/ncbitaxon.owl>
4 WHERE {
5   VALUES ?ncbitaxonURI {<http://purl.obolibrary.org/obo/NCBITaxon_5296>}
6   GRAPH <http://ns.inria.fr/d2kab/graph/wheatgenomicsslkg> {
7     ?a1 a oa:Annotation ;
8       oa:hasTarget [ oa:hasSource ?source1 ] ;
9       oa:hasBody [ a d2kab:Gene ; skos:prefLabel ?geneName ] .
10    ?source1 frbr:partOf+ ?paper .
11    ?a3 a oa:Annotation ;
12      oa:hasTarget [ oa:hasSource ?source2 ] ;
13      oa:hasBody ?ncbiTaxonURI .
14    ?source2 frbr:partOf+ ?paper .
15    ?paper a fabio:ResearchPaper ; dct:title ?titleURI .
16    ?titleURI rdf:value ?title . }
17   GRAPH <http://purl.obolibrary.org/obo/ncbitaxon/ncbitaxon.owl> {
18     ?ncbiTaxonURI rdfs:subClassOf* ?ncbitaxonURI ; rdfs:label ?ncbiTaxon . }
19 } 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).

539 4.2.2 SPARQL Queries Implementing CQs on the PHB Corpus

540 The SPARQL queries implementing CQ on the PHB corpus can be executed at the endpoint
541 <http://ontology.inrae.fr/bsv/sparql>. Since not all FCU crop concepts have English labels, the
542 results of the queries may vary depending on the label language. In the following SPARQL
543 queries, only French labels are requested. A Jupyter Notebook of these SPARQL queries is
544 available on our Github repository ³².

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

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

553 **CQ7** : The SPARQL query presented in Listing 9 implements CQ7 and retrieves the number

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

```

1 SELECT ?phb ?cropName ?xpath
2 WHERE {
3   SERVICE <http://ontology.inrae.fr/frenchcropusage/sparql> {
4     ?body a skos:Concept ; skos:prefLabel ?cropName . FILTER (LANG(?cropName)='fr') . }
5   ?phb a d2kab_inrae:Bulletin ; dul:isRealizedBy ?phb_html .
6   ?phb_html dce:format "text/html" .
7   ?rs a oa:ResourceSelection ; oa:hasSelector ?sel ; oa:hasSource ?phb_html .
8   ?aa a oa:Annotation ; oa:hasTarget ?rs ; oa:hasBody ?body .
9   ?sel a oa:XPathSelector ; rdf:value ?xpath . FILTER contains(?xpath, "/H1[") .
10 } LIMIT 50

```

Listing 7: SPARQL query implementing CQ6 and retrieving the crop names appearing in H1 sections of the HTML versions of PHBs

```

1 SELECT ?cropName (count(?cropName) AS ?nb)
2 WHERE {
3   <http://ontology.inrae.fr/bsv/resources/Q16994/2019/bsv_viti_13_27_06_2019_cle07f426>
4     a d2kab_inrae:Bulletin ; dul:isRealizedBy ?phb_html .
5   ?phb_html dce:format "text/html" .
6   ?rs a oa:ResourceSelection ; oa:hasSource ?phb_html .
7   ?aa a oa:Annotation ; oa:hasTarget ?rs ; oa:hasBody ?body .
8   SERVICE <http://ontology.inrae.fr/frenchcropusage/sparql> {
9     ?body a skos:Concept ; skos:prefLabel ?cropName . FILTER (LANG(?cropName)='fr') }
10 }
11 GROUP BY ?cropName
12 ORDER BY DESC(?nb)

```

Listing 8: SPARQL query implementing CQ6 and computing the number of times that a crop name appears in PHB-KG

554 of times that each crop name is mentioned in the subset of bulletins for the French region *Pays*
555 *de la Loire*, ordered by descending order. It estimates the most cultivated crops in this region
556 during the whole time period. This query could also include a specific time period to observe
557 the evolution of cultivated crop in the region. The results show that grape, cabbage, leek and
558 carrot are the most cultivated crops in *Pays de la Loire*. These are more precise results than
559 those from our previous work [RBP⁺17] in 2017 indicating that this region growth field crops,
560 vegetables and fruits.

561 **CQ8** : The SPARQL query presented in Listing 10 implements CQ8 and retrieves couples
562 of annotations, one for a crop name and one for a development stage, that are localized in
563 the same HTML element of a PHB. This query then estimates that the development stage is
564 applicable to the crop. Thus one can deduce that at the publication date of the PHB the crop
565 has reached the development stage in the region the PHB is relative to. The query retrieves

```

1 SELECT ?cropName (count(?cropName) AS ?nb)
2 WHERE {
3   SERVICE <http://ontology.inrae.fr/frenchcropusage/sparql> {
4     ?body a skos:Concept ; skos:prefLabel ?cropName . FILTER (LANG(?cropName)='fr') }
5   ?phb a d2kab_inrae:Bulletin ;
6     dct:spatial wikidata:Q16994 ;
7     dul:isRealizedBy ?phb_html .
8   ?phb_html dce:format "text/html" .
9   ?rs a oa:ResourceSelection ; oa:hasSource ?phb_html .
10  ?aa a oa:Annotation ; oa:hasTarget ?rs ; oa:hasBody ?body .
11 }
12 GROUP BY ?cropName
13 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.

566 1304 HTML elements that contain both an annotation of FCU crop concepts and of PPDO
567 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.

```

1 SELECT ?phb ?cropName ?devtName ?xpt
2 WHERE {
3   SERVICE <http://ontology.inrae.fr/ppdo/sparql> {
4     ?body_devt a owl:NamedIndividual ;
5       skos:inScheme <http://ontology.inrae.fr/ppdo/ontology/bbch_globalScale> ;
6       skos:prefLabel ?devtName . FILTER (LANG(?devtName)='fr') }
7   ?phb a d2kab_inrae:Bulletin ; dul:isRealizedBy ?phb_html .
8   ?phb_html dce:format "text/html" .
9   ?rs1 a oa:ResourceSelection ; oa:hasSource ?phb_html ; oa:hasSelector ?sel1 .
10  ?aa1 a oa:Annotation ; oa:hasTarget ?rs1 ; oa:hasBody ?body_fcu .
11  ?sel1 a oa:XPathSelector ; rdf:value ?xpt .
12  ?rs2 a oa:ResourceSelection ; oa:hasSource ?phb_html ; oa:hasSelector ?sel2 .
13  ?sel2 a oa:XPathSelector ; rdf:value ?xpt .
14  ?aa2 a oa:Annotation ; oa:hasTarget ?rs2 ; oa:hasBody ?body_devt .
15  SERVICE <http://ontology.inrae.fr/frenchcropusage/sparql> {
16    ?body_fcu a skos:Concept ; skos:prefLabel ?cropName . FILTER (LANG(?cropName)='fr') }
17 } 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.

568
569 In the PHB example of Figure 2, the crop name and the development stage are mentioned in
570 distinct HTML elements. The crop name is mentioned before the development stage. Another
571 implementation consists in retrieving the crop name and the development stage in a window
572 of a fixed number of characters. The following query retrieves the couples of annotations, one

573 for the crop name and one for the development stage, that appear in the characters window of
 574 1000 characters in the PHB example. This alternative implementation of CQ8 is presented in
 575 Listing 11.

```

1 SELECT ?phb ?pos1 ?cropName ?pos2 ?devtName
2 WHERE {
3   SERVICE <http://ontology.inrae.fr/ppdo/sparql> {
4     ?body_devt a owl:NamedIndividual ;
5       skos:inScheme <http://ontology.inrae.fr/ppdo/ontology/bbch_globalScale> ;
6       skos:prefLabel ?devtName . FILTER (LANG(?devtName)='fr') }
7   ?phb a d2kab_inrae:Bulletin ; dul:isRealizedBy ?phb_html .
8   ?phb_html dce:format "text/html" .
9   ?rs1 a oa:ResourceSelection ; oa:hasSource ?phb_html ; oa:hasSelector ?sel1 .
10  ?aa1 a oa:Annotation ; oa:hasTarget ?rs1 ; oa:hasBody ?body_fcu .
11  ?sel1 a oa:TextPositionSelector ; oa:start ?pos1 .
12  ?rs2 a oa:ResourceSelection ; oa:hasSource ?phb_html ; oa:hasSelector ?sel2 .
13  ?sel2 a oa:TextPositionSelector ; oa:start ?pos2 .
14  FILTER (abs(?pos2-?pos1) < 1000)
15  ?aa2 a oa:Annotation ; oa:hasTarget ?rs2 ; oa:hasBody ?body_devt .
16  SERVICE <http://ontology.inrae.fr/frenchcropusage/sparql> {
17    ?body_fcu a skos:Concept ; skos:prefLabel ?cropName . FILTER (LANG(?cropName)='fr') }
18 } LIMIT 20

```

Listing 11: Another SPARQL implementation of CQ8 retrieving couples of annotations located within a window of a fixed number of characters.

576 4.2.3 Combined Exploitation of Knowledge Graphs

577 Using federated queries, scientists can jointly exploit several KGs. In the following, we present
 578 an example combined exploitation of WheatGenomicsSLKG and RiceGenomicsSLKG and an
 579 example combined exploitation of WheatGenomicsSLKG and PHB KG.³³

580 Combined Exploitation of WheatGenomicsSLKG and RiceGenomicsSLKG CQ5

581 requires to use both wheat and rice KGs to search a correlation between gene expression and
 582 disease resistance. The aim is to search for wheat and rice genes co-occurring with the same
 583 taxon of a pathogen (or a more specific taxon) to identify candidate orthologous genes in wheat
 584 and rice genomes. The SPARQL query presented in Listing 12 implements this competency
 585 question. Starting with the *Magnaporthe oryzae* URI³⁴ or its upper parent³⁵, we retrieve wheat

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

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

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

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

```
1 SELECT distinct ?paper ?title (GROUP_CONCAT(distinct ?geneName; SEPARATOR="-") as ?genes)
2   ?ncbiTaxon
3 WHERE {
4   ?a1 a oa:Annotation ;
5     oa:hasTarget [ oa:hasSource ?source1 ] ;
6     oa:hasBody [ a d2kab:Gene; skos:prefLabel ?geneName ] .
7   ?source1 frbr:partOf+ ?paper .
8   ?a2 a oa:Annotation;
9     oa:hasTarget [ oa:hasSource ?source2 ] ;
10    oa:hasBody ?ncbitaxonURI .
11   ?source2 frbr:partOf+ ?paper .
12   ?paper a fabio:ResearchPaper ; dct:title ?titleURI .
13   ?titleURI rdf:value ?title .
14   GRAPH <http://purl.obolibrary.org/obo/ncbitaxon/ncbitaxon.owl> {
15     ?ncbitaxonURI rdfs:subClassOf* <http://purl.obolibrary.org/obo/NCBITaxon_639021> ;
16     rdfs:label ?ncbiTaxon . }
17 }
```

Listing 12: SPARQL federated query implementing CQ5 and allowing the combined exploitation of WheatGenomicsSLKG and RiceGenomicsSLKG to retrieve orthologous genes.

588 **Combined Exploitation of WheatGenomicsSLKG and PHB KG** The SPARQL query
589 presented in Listing 13 implements CQ9: it enables to retrieve publications in PubMed and PHB
590 bulletins corpora mentioning the same taxon (*Triticum aestivum*). As each corpus uses different
591 semantic resources to annotate taxon entities (NCBI taxonomy in WheatGenomicsSLKG, and
592 FCU thesaurus in PHB KG), the query exploits a third KG, TAXREF-LD³⁶ [MGTFZ17] to
593 retrieve the alignments between NCBI classes and FCU concepts (line 20). Alignments between
594 FCU concepts and TAXREF-LD classes were generated automatically based on the *Official*
595 *Catalogue of Species and Varieties of Cultivated Crops*, which is denoted by the alignment
596 predicate `taxref:candidateAlignment_geves`. This way, the query retrieves that taxon <http://taxref.mnhn.fr/lod/taxon/127692>
597 is aligned with FCU concepts `fcu:Bles_tendres_hiver`³⁷
598 and `fcu:Bles_tendres_printemps`³⁸. Note that this example query is meant to be executed
599 on the WheatGenomicsSLKG SPARQL endpoint, invoking the two other SPARQL endpoints

³⁶TAXREF-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>

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

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

600 via SERVICE clauses. This illustrates the fact that publishing KGs according to FAIR design
 601 and publication principles allows to achieve the important goal of querying uniformly several
 602 interoperable knowledge graphs.

```

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

```

Listing 13: SPARQL federated query implementing CQ9 and allowing the combined exploitation of WheatGenomicsSLKG and PHB KGs

603 5 Discussion and Lessons Learned

604 We produced three different knowledge graphs compliant with the semantic model presented in
 605 Section 3.2 and representing the annotations of three text corpora with automatically extracted
 606 NEs. The proposed semantic model is independent from the NLP tools used for NERL and
 607 may be reused for representing annotations extracted using any NLP tools. The annotations
 608 rely on the same understanding of the OA model. Each annotation links one mention in a
 609 document to one domain-specific entity defined in external semantic resources. As illustrated
 610 in Figure 6, two different annotations may annotate the same piece of text, hence sharing the
 611 same target resource (an entity mention in the PHB document) while having two different
 612 bodies (two different domain concepts).

613 Furthermore, as presented in Section 3.2, the OA model provides different types of se-
 614 lectors which are generic enough to fulfill different needs considering the representation of

615 the source resource. To represent the annotations of the PubMed and PHB corpora, we
616 used three different selectors proposed by OA to precisely locate an entity mention in a text:
617 `oa:TextQuoteSelector`, `oa:TextPositionSelector`, and `oa:XPathSelector`. The first two
618 selectors are used in both corpora. In particular the `oa:TextPositionSelector` selector locates
619 the entity mention within this part of the document. `oa:XPathSelector` is used for the PHB
620 corpus to identify the HTML element in which an entity mention was recognized. Note that
621 the document structures are not represented in the same way in both corpora. In the PubMed
622 corpus, each document is identified by a URI which corresponds to its entry in the PubMed
623 repository. Entity mentions may be extracted from the title, abstract or abstract's sub-parts,
624 each having its own URI and being linked by the `frbr:partOf` property (Figure 7). In the
625 PHB corpus, an HTML document is identified by a URI that is the source of the annotation.
626 This work illustrates that OA selectors are sufficiently broad to locate entities mentions in a
627 variety of situations.

628 The coverage scope of our model could be extended by identifying complementary vocab-
629 ularies. In particular, information such as frequency, lexical and morphological characteristics
630 that can be drawn from text corpora and semantic resources can be added to our model by
631 reusing terms from the FrAC vocabulary, an OntoLex module for Frequency, Attestation and
632 Corpus information (FrAC) [CIDD⁺20]. FrAC allows to model absolute frequencies of a given
633 lexical entity (how many times an element of a semantic resource is recognized in the text,
634 e.g., as shown in CQ5 bis) which is a recurrent need. Figure 10 presents an RDF graph that
635 links six instances of class `oa:Annotations` to one instance of the `frac:CorpusFrequency`
636 class using the `prov:wasDerivedFrom` property. Those annotations are about the FCU con-
637 cept "grapevine" recognized in the PHB example of Figure 2. Considering that a document is
638 a corpus of one element, the instance of class `frac:CorpusFrequency` represents the number
639 of times that the given concept (grapevine) is recognized in the PHB document.

640 6 Conclusions and Future Works

641 In this paper, we presented the results of a research work to support scientists with method-
642 ologies to standardize and share domain knowledge extracted from texts according to FAIR

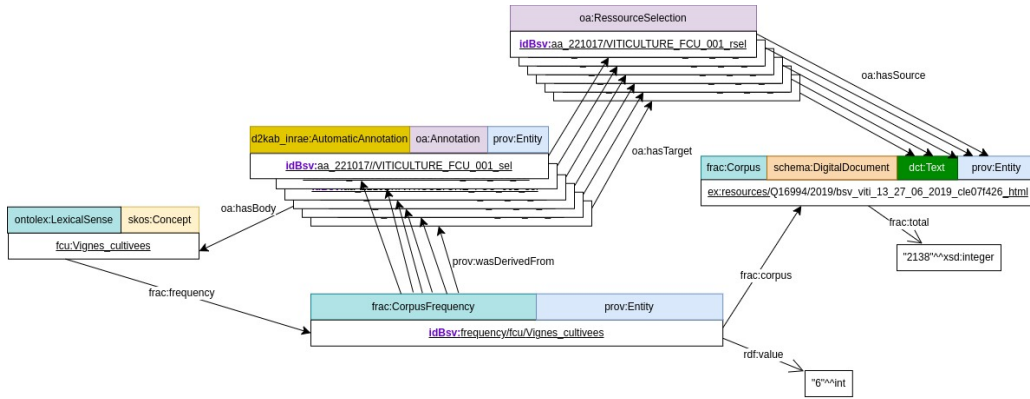


Figure 10: Frequency Modelling in PHB graph based on FRAC, OA and PROV vocabularies.

643 principles. We rely on Semantic Web models and technologies to build domain-specific KGs
 644 that allow domain experts to explore and retrieve information from the annotated corpora and
 645 deduce new domain knowledge. Our approach relies on the formalization of a unified semantic
 646 data model to describe, structure and integrate annotations using NE automatically recog-
 647 nized from texts in the agricultural domain. NE entities normalization and linking is based
 648 on semantic resources widely adopted in Agriculture (for phenotypes, traits, taxa, cultivated
 649 varieties). The core part of this model is based on the W3C Web Annotation Ontology (OA)
 650 which has been complemented by eight different vocabularies to describe documents metadata
 651 and provenance information. We used this model to construct three different knowledge graphs
 652 from three distinct agricultural corpora using a mapping-based transformation pipeline. The
 653 relevance of the semantic model was validated by implementing a set of competency questions
 654 with SPARQL queries which reflect how the KGs can be queried to retrieve co-occurrence of
 655 NE in texts. The proposed semantic model and generation pipeline are generic enough to be
 656 reused to build new KGs in different research domains in order to enable scientists explore their
 657 scientific literature.

658 As future works, we want to investigate the extraction of relations between recognized
 659 NE. Several competency questions involve retrieving entities that appear in the same context
 660 within texts. They could be refined by precisising the relationship between the entities. As
 661 relation extraction strongly depends on the accuracy of the entity recognition task, an important
 662 first step for the PHB knowledge graph will focus on the improvement of the accuracy of NE
 663 annotations which is not always satisfying so far.

664 On another note, we plan to construct and publish richly annotated gold standard datasets
665 based on the three corpora. This will require considerable efforts of domain experts to define
666 guidelines and samples for NE annotations. Gold-standard datasets can be used to train and
667 evaluate natural language processing (NLP) approaches, such as specialized NE recognition,
668 relation extraction and entity linking.

669 Funding

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

672 Material availability

673 The SKOS version of the WTO used in this study can be queried: [http://d2kab.i3s.unice.fr/
674 sparql](http://d2kab.i3s.unice.fr/sparql).

675 The NCBITaxon used for this study can be found in the OBO Foundry repository: [https:
676 //obofoundry.org/ontology/ncbitaxon.html](https://obofoundry.org/ontology/ncbitaxon.html).

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

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

683 The SPARQL endpoint of the KG on wheat and rice genomics scientific literature: [http:
684 //d2kab.i3s.unice.fr/sparql](http://d2kab.i3s.unice.fr/sparql).

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

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

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

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

693 References

- 694 [CIDD⁺20] Christian Chiarcos, Maxim Ionov, Jesse de Does, Katrien Depuydt, Fahad Khan,
695 Sander Stolk, Thierry Declerck, and John Philip McCrae. Modelling frequency
696 and attestations for ontolox-lemon. In Proceedings of the 2020 Globalex Workshop
697 on Linked Lexicography, pages 1–9, 2020.
- 698 [DFMd19] Brett Drury, Robson Fernandes, Maria-Fernanda Moura, and Alneu de An-
699 drade Lopes. A survey of semantic web technology for agriculture. Information
700 Processing in Agriculture, 6(4):487–501, 2019.
- 701 [GF95] Michael Grüninger and Mark S. Fox. The Role of Competency Questions in
702 Enterprise Engineering, pages 22–31. Springer US, Boston, MA, 1995.
- 703 [JTA⁺18] Clément Jonquet, Anne Toulet, Elizabeth Arnaud, Sophie Aubin, Esther Dzalé
704 Yeumo, Vincent Emonet, John Graybeal, Marie-Angélique Laporte, Mark A.
705 Musen, Valeria Pesce, and Pierre Larmande. Agroportal: A vocabulary and
706 ontology repository for agronomy. Computers and Electronics in Agriculture,
707 144:126–143, 2018.
- 708 [KCD⁺22] Anas Fahad Khan, Christian Chiarcos, Thierry Declerck, Daniela Gifu,
709 Elena González-Blanco García, Jorge Gracia, Maxim Ionov, Penny Labropoulou,
710 Francesco Mambrini, John P. McCrae, Émilie Pagé-Perron, Marco Passarotti,
711 Salvador Ros Muñoz, and Ciprian-Octavian Truica. When linguistics meets web
712 technologies. recent advances in modelling linguistic linked data. Semantic Web,
713 13(6):987–1050, 2022.
- 714 [KY06] Nori Kurata and Yukiko Yamazaki. Oryzabase. An Integrated Biological and
715 Genome Information Database for Rice. Plant Physiol., 140(1):12, January 2006.

- 716 [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.
- 717
- 718
- 719
- 720 [Mei18] Uwe Meier. Growth stages of mono- and dicotyledonous plants: BBCH Monograph. Open Agrar Repository, 2018.
- 721
- 722 [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.
- 723
- 724
- 725
- 726 [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.
- 727
- 728 [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. Winkler. 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.
- 729
- 730
- 731
- 732
- 733
- 734 [MGTFZ17] Franck Michel, Olivier Gargominy, Sandrine Terceirie, 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>.
- 735
- 736
- 737
- 738
- 739
- 740 [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.
- 741
- 742

- 743 [NBV⁺14] Claire Nédellec, Robert Bossy, Dialekti Valsamou, Marion Ranoux, Wiktorina Go-
744 lik, and Pierre Sourdille. Information extraction from bibliography for marker-
745 assisted selection in wheat. In Proceedings of Research Conference on Metadata
746 and Semantics Research - MTSR 2014, pages pp. 301–313, 2014.
- 747 [NIBS20] Claire Nédellec, Liliana L. Ibanescu, Robert Bossy, and Pierre Sourdille.
748 WTO, an ontology for wheat traits and phenotypes in scientific publications.
749 Genomics & Informatics, 18(2), 2020.
- 750 [Per16] Silvio Peroni. SAMOD: an agile methodology for the development of ontologies.
751 In Mauro Dragoni, Maréa Poveda-Villalón, and Ernesto Jimenez-Ruiz, editors,
752 OWL: Experiences and Directions – Reasoner Evaluation, pages 55–69. Springer,
753 2016.
- 754 [PG08] Valentina Presutti and Aldo Gangemi. Content ontology design patterns as prac-
755 tical building blocks for web ontologies. In International conference on conceptual
756 modeling, pages 128–141. Springer, 2008.
- 757 [PS12] Silvio Peroni and David Shotton. FaBiO and CiTO: Ontologies for describing
758 bibliographic resources and citations. Journal of Web Semantics, 17:33–43, 2012.
- 759 [RBP⁺17] Catherine Roussey, Stephan Bernard, François Pinet, Xavier Reboud, Vincent
760 Cellier, Ivan Sivadon, Danièle Simonneau, and Anne-Laure Bourigault. A method-
761 ology for the publication of agricultural alert bulletins as lod. Computers and
762 Electronics in Agriculture, 142:632–650, 2017.
- 763 [RDA⁺21] Catherine Roussey, Xavier Delpuech, Florence Amardeilh, Stephan Bernard, and
764 Clement Jonquet. Semantic description of plant phenological development stages,
765 starting with grapevine. In Emmanouel Garoufallou and María-Antonia Ovalle-
766 Perandones, editors, Metadata and Semantic Research, pages 257–268, Cham,
767 2021. Springer International Publishing.
- 768 [SCD⁺20] Conrad L. Schoch, Stacy Ciufu, Mikhail Domrachev, Carol L. Hotton, Sivakumar
769 Kannan, Rogneda Khovanskaya, Detlef D. Leipe, Richard McVeigh, Kathleen

- 770 O’Neill, Barbara Robbertse, Shobha Sharma, Vladimir Soussov, John P. Sullivan,
771 Lu Sun, Seán Turner, and Ilene Karsch-Mizrachi. Ncbi taxonomy: a comprehen-
772 sive update on curation, resources and tools. Database : the journal of biological
773 databases and curation, 2020, 2020.
- 774 [SCY17] Robert Sanderson, Paolo Ciccarese, and Benjamin Young. Web annotation ontol-
775 ogy. Technical report, W3C, 2017.
- 776 [The22] The UniProt Consortium . UniProt: the Universal Protein Knowledgebase in
777 2023. Nucleic Acids Research, 51(D1):D523–D531, 11 2022.
- 778 [UG96] Mike Uschold and Michael Gruninger. Ontologies: Principles, methods and ap-
779 plications. The Knowledge Engineering Review, 11(2):93–136, 1996.
- 780 [VTNH⁺18] Aravind Venkatesan, Gildas Tagny Ngompe, Nordine El Hassouni, Imene Chentli,
781 Valentin Guignon, Clement Jonquet, Manuel Ruiz, and Pierre Larmande. Agro-
782 nomic Linked Data (AgroLD): A knowledge-based system to enable integrative
783 biology in agronomy. PLOS ONE, 13(11):1–17, 2018.
- 784 [WDA⁺16] Mark Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gaby Apple-
785 ton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Olavo
786 Bonino da Silva Santos, Philip Bourne, Jildau Bouwman, Anthony Brookes, Tim
787 Clark, Merce Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris Evelo,
788 Richard Finkers, and Barend Mons. The fair guiding principles for scientific data
789 management and stewardship. Scientific Data, 3, 03 2016.
- 790 [WSM⁺20] Leon Weber, Mario Sanger, Jannes Munchmeyer, Maryam Habibi, Ulf Leser,
791 and Alan Akbik. HunFlair: An Easy-to-Use Tool for State-of-the-Art Biomed-
792 ical Named Entity Recognition. arXiv:2008.07347 [cs], August 2020. arXiv:
793 2008.07347.
- 794 [YBC⁺22] Eric Yao, Victoria C Blake, Laurel Cooper, Charlene P Wight, Steve Michel,
795 H Busra Cagirici, Gerard R Lazo, Clay L Birkett, David J Waring, Jean-Luc Jan-
796 nink, Ian Holmes, Amanda J Waters, David P Eickholt, and Taner Z Sen. Grain-

797

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

798

2022, 05 2022. baac034.