The Wheat and Rice Genomics Scientific Literature Knowledge Graphs

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

[1]Université Claude Bernard Lyon 1, CNRS, LIRIS (UMR 5205), France
[2]Université Côte d'Azur, Inria, CNRS, I3S (UMR 7271), France
[4]MaIAGE, INRAE, Université Paris-Saclay, 78350 Jouy-en-Josas, France
[5]DIADE, IRD, CIRAD, Univ. Montpellier, Montpellier, France

Abstract

This paper presents a generic semantic model to describe, structure, and integrate the named entities automatically extracted from scientific texts, represented as annotations. This model has been used to construct knowledge graphs from two distinct agricultural corpora consisting of PubMed scientific publications on wheat and rice genetics. The named entities to be recognized are genes, phenotypes, traits, genetic markers, and taxa. For both corpora, named entities were automatically extracted using natural language processing tools. The RDF model was populated using a mapping-based transformation pipeline implemented with the Morph-xR2RML tool which takes CSV files as input. The resulting RDF knowledge graphs are deployed and query-able through dedicated web applications.

Keywords

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

This paper presents a methodology for constructing two domain-specific knowledge graphs by extracting relevant entities from textual scientific corpora and organizing them in a structured and meaningful way. The methodology uses semantic Web technologies, which involves the re-use of shared RDF-based standard vocabularies.

The MaIAGE research group[1] collected 8,496 scientific articles published between 1974 and 2021, related to wheat selection. We used the AlvisNLP [1] workflow to identify named entities (NE) and the relationships between wheat varieties and phenotypes. In total, 88,880 mentions of 4,318 distinct named entities were identified from PubMed abstracts and titles. Similarly, the DIADE research group[2] collected 17,058 scientific articles published between 1951 and 2021 from the Oryzabase database [2] which provides manually checked PubMed entries related to rice genomics. We used the HunFLAIR NER tagger [3] to extract NEs in the title and the abstract of the articles. In total, 351,003 mentions of 63,591 distinct NEs were identified.

Both pipelines distinguish between NE mentions that refer to genes, genetic markers, traits, phenotypes, taxa, and cultivar entities mentioned in the title and abstract of publications. When possible, these NEs were linked with existing semantic resources. The wheat phenotype and trait mentions are linked to classes in the Wheat Trait Ontology[3] (WTO), and taxon mentions are linked to NCBI[4] taxonomy classes.

In both graphs the core part of the data model is based on the W3C Web Annotation Ontology (OA) which has been complemented with different vocabularies to describe documents metadata described in Yacoubi et al. [4]. The construction pipeline involves two main steps. Firstly, we use a SPARQL micro-service [5] to query Pubmed's Web API and translate the articles' metadata (including the title and abstract) into RDF[5]. Secondly, AlvisNLP [1] and HunFLAIR [3] are used to extract and link the named entities mentioned in the titles and abstracts. The output consists of CSV files that are translated to RDF using Morph-xR2RML [6].

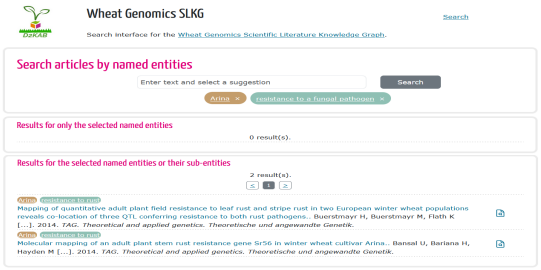
SWAT4HCLS 2025: 16th International Conference on Semantic Web Applications and Tools for Health Care and Life Sciences 2025
*Corresponding author.
†These authors contributed equally.

[1]https://maiage.inrae.fr/en
[2]http://diade.ird.fr/
[3]https://doi.org/10.57745/TAQQFZ
[4]https://www.ncbi.nlm.nih.gov/taxonomy
[5]https://sparql-micro-services.org/service/pubmed/getArticleByPMId_sd/

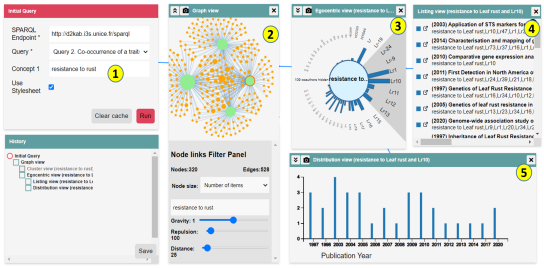
We developed two web applications to enable researchers to explore the wheat and rice knowledge graphs. The first interface is depicted in Figure 1a. In this text-based search, a user chooses some NEs of interest (taking into account the various labels and languages). The results list all the articles that contain these NEs or any of their sub-entities (sub-concepts, sub-classes). In the example, selecting *resistance to fungal pathogen* returns the articles mentioning the *resistance to rust* trait which is a sub-class of *resistance to fungal pathogen* in WTO.

The second interface, depicted in Figure 1b, provides domain experts with predefined queries (1) to explore the data through complementary visualization techniques. The exploration starts with a graph view (2) where nodes represent NEs (here traits or genes) linked through the scientific publications where they co-occur. The egocentric view (3) focused on the *resistance to leaf rust* shows the different genes co-occurring with this trait, and the number of publications where they co-occur. The listing view (4) then shows the list of publications concerned by this co-occurrence, together with their time distribution (5).

The code and material for building both graphs is available publicly under open licenses[6].



(a) Search for articles mentioning the Arina variety and the fungal pathogen resistance trait or any of its sub-concepts.



(b) Visual exploration using LDViz to discover genes mentioned proximal to the *rust resistance* trait in scientific literature.

Figure 1: Visual exploration of the Wheat Genomics Scientific Literature Knowledge Graph

Declaration on Generative AI

The authors have not employed any Generative AI tools.

References

[1] M. Ba, R. Bossy, Interoperability of corpus processing work-flow engines: the case of alvisnlp/ml in openminted, in: Proceedings of the INTEROP 2016 Workshop, organised with LREC 2016, Portorož, Slovenia, 2016, pp. pp. 15–18.

[2] N. Kurata, Y. Yamazaki, Oryzabase. An Integrated Biological and Genome Information Database for Rice, Plant Physiol. 140 (2006) 12. doi:10.1104/pp.105.063008.

[3] L. Weber, M. Sänger, J. Münchmeyer, M. Habibi, U. Leser, A. Akbik, HunFlair: An Easy-to-Use Tool for State-of-the-Art Biomedical Named Entity Recognition, Bioinformatics 37 (2021) 2792–2794. doi:10.1093/bioinformatics/btab042.

[4] N. Yacoubi Ayadi, S. Bernard, R. Bossy, M. Courtin, B. G. Happi Happi, P. Larmande, F. Michel, C. Nédellec, C. Roussey, C. Faron, A unified approach to publish semantic annotations of agricultural documents as knowledge graphs, Smart Agricultural Technology 8 (2024) 100484. doi:10.1016/j.atech.2024.100484.

[5] F. Michel, C. Faron Zucker, O. Gargominy, F. Gandon, Integration of Web APIs and Linked Data Using SPARQL Micro-Services - Application to Biodiversity Use Cases, Information 9 (2018). URL: https://hal.science/hal-01947589. doi:10.3390/info9120310.

[6] https://github.com/Wimmics/wheatgenomicsslkg, https://github.com/ANR-DIG-AI/RiceGenomicsSLKG

[6] F. Michel, L. Djimenou, C. Faron-Zucker, J. Montagnat, Translation of Relational and Non-Relational Databases into RDF with xR2RML, in: Proceeding of the 11th WebIST conference, Lisbon, Portugal, 2015, pp. 443–454. doi:10.5220/0005448304430454.