PotatoG-DKB: a potato gene-disease knowledge base mined from biological literature

Congjiao Xie[1,2], Jing Gao[1,2,3], Junjie Chen[1,2] and Xuyang Zhao[1]

[1] College of Computer and Information Engineering, Inner Mongolia Agricultural University, Hohhot, Inner Mongolia, China
[2] Inner Mongolia Autonomous Region Key Laboratory of Big Data Research and Application for Agriculture and Animal Husbandry, Hohhot, Inner Mongolia, China
[3] Inner Mongolia Autonomous Region Government Service and Data Management Bureau, Hohhot, Inner Mongolia, China

Submitted 18 April 2024
Accepted 9 September 2024
Published 3 October 2024

Corresponding author
Jing Gao, gaojing@imau.edu.cn

Academic editor
Yuriy Orlov

Additional Information and Declarations can be found on page 16

DOI 10.7717/peerj.18202

ABSTRACT

Background: Potato is the fourth largest food crop in the world, but potato cultivation faces serious threats from various diseases and pests. Despite significant advancements in research on potato disease resistance, these findings are scattered across numerous publications. For researchers, obtaining relevant knowledge by reading and organizing a large body of literature is a time-consuming and labor-intensive process. Therefore, systematically extracting and organizing the relationships between potato genes and diseases from the literature to establish a potato gene-disease knowledge base is particularly important. Unfortunately, there is currently no such gene-disease knowledge base available.

Methods: In this study, we constructed a Potato Gene-Disease Knowledge Base (PotatoG-DKB) using natural language processing techniques and large language models. We used PubMed as the data source and obtained 2,906 article abstracts related to potato biology, extracted entities and relationships between potato genes and related disease, and stored them in a Neo4j database. Using web technology, we also constructed the Potato Gene-Disease Knowledge Portal (PotatoG-DKP), an interactive visualization platform.

Results: PotatoG-DKB encompasses 22 entity types (such as genes, diseases, species, *etc.*) of 5,206 nodes and 9,443 edges between entities (for example, gene-disease, pathogen-disease, *etc.*). PotatoG-DKP can intuitively display associative relationships extracted from literature and is a powerful assistant for potato biologists and breeders to understand potato pathogenesis and disease resistance. More details about PotatoG-DKP can be obtained at https://www.potatogd.com.cn/.

Subjects Agricultural Science, Bioinformatics, Data Mining and Machine Learning
Keywords Knowledge base, Literature mining, Potato, Large language model, Disease

INTRODUCTION

Potato (*Solanum tuberosum* L.) is a member of the Solanaceae family. Because of its high nutritional and economic value, potato is the fourth largest food crop worldwide, following wheat, maize, and rice (*Kondhare, Natarajan & Banerjee, 2020*). In 2022, the Food and Agriculture Organization of the United Nations (FAO) reported that global potato

How to cite this article Xie C, Gao J, Chen J, Zhao X. 2024. PotatoG-DKB: a potato gene-disease knowledge base mined from biological literature. **PeerJ 12:e18202 DOI 10.7717/peerj.18202**

production surpassed 470 million tons, and as a major potato producing country, China's potato production exceeded 190 million tons (*FAO, 2022*). However, potatoes suffer from various diseases and pest infestations, seriously impairing tuber quality and yield (*Yuen, 2021*; *Enciso-Rodriguez et al., 2018*). Fungicides currently serve as an essential shield against diseases and pests (*Hernández, 2014*; *Tiwari et al., 2021*); unfortunately, their pervasive application poses both environmental and health risks (*Majeed et al., 2017a*, *2017b*). As crop genomes continue to emerge, an increasing number of biologists are investigating gene-disease relationships at the gene level and publishing their findings.

A knowledge base facilitates the integration and efficient retrieval of information. Currently, there are three main methods for constructing a knowledge base: one based on genomic sequences, another based on ontology, and the third based on literature mining. Genomic sequencing provides detailed information on the genetic sequences of organisms, enabling a comprehensive understanding of gene structure and function. However, a knowledge base built with genomic sequencing lacks correlations between findings. The ontology-based approach typically relies on predefined ontologies, which include formalized descriptions of concepts, properties, and relationships. In contrast, literature mining extracts key information from a vast amount of scientific literature, establishing relationships between different pieces of knowledge. To date, the PubMed database houses more than 12,787 scientific articles related to *Solanum tuberosum*. A potato-related knowledge network based on literature mining can aid researchers in quickly finding relevant studies and data, thereby promoting the discovery of new biological mechanisms.

Multiple potato knowledge bases have been created, as shown in Table 1. The Spud DB (*Hirsch et al., 2014*), based on genome sequences and associated annotation datasets, offers phenotypic and genotypic data, functional annotation, gene ontology (GO), and BLAST databases from 250 potato clones. The SolRgene (*Vleeshouwers et al., 2011*) contains the R gene and R gene homologues of *Solanum* petota; the resistance of *Solanum* petota to late blight was analyzed through high-throughput disease experiments and field tests under various laboratory conditions. The PoMaMo database (*Meyer, Nagel & Gebhardt, 2005*) is a comprehensive archive of potato genome data. *Choe et al. (2018)* developed a comparative analysis tool and a genome viewer using web technology to identify homologous genes based on collinearity and sequence similarity among tomato, pepper, and potato species. Numerous plant genome databases exist that are not specific to potato, such as Ensembl plants (*Bolser et al., 2017*) with around 70 plant genomes, Crop Ontologies (*Shrestha et al., 2012*), which provides ontology-based descriptors for crop traits and standard variables for over 20 crops, and the Planteome database (*Cooper et al., 2018*), which provides comprehensive resources for reference ontology, plant genomics, and phenotypic genomics. Some knowledge bases also use the literature mining method, including the *Solanum tuberosum* knowledge base (*Ivanisenko et al., 2018*), which is a knowledge base of molecular genetic regulation and metabolic pathways. *Singh et al. (2021)* took the flesh color of potatoes as the main trait and trained an NLP model to recognize relevant biological entities and relationships (genes, proteins, metabolites, and traits) in a manually annotated *corpus* of 34 full-text potato articles. The International Potato Centre

Table 1 **Potato-related knowledge bases.** For each potato-related knowledge base introduced, this table includes the name of the knowledge base, the methods used for its construction, the species covered by the knowledge base, the primary data content stored in the knowledge base, the scale of the knowledge base, and the URL.

Knowledge base	Method	Species	Content	Scale	Source URL
Spud DB (*Hirsch et al., 2014*)	Genomic sequence	Potato	Potato genome sequence and associated annotation datasets	250 potato clones	http://spuddb.uga.edu/
SolRgene (*Vleeshouwers et al., 2011*)	Genomic sequence	Solanum	R genes to *Phytophthora infestans*	1,062 accessions, and 12,584 genotypes	https://www.plantbreeding.wur.nl/SolRgenes/
PoMaMo (*Meyer, Nagel & Gebhardt, 2005*)	Genomic sequence	Potato	Potato genome data	Sequence data and molecular maps of all 12 potato chromosomes with about 1,000 mapped elements	https://www.gabipd.org/projects/Pomamo/
Ensembl plants (*Bolser et al., 2017*)	Genomic sequence	Plant	Genome data	39 sequenced plant species	https://plants.ensembl.org/index.html
Crop ontology (*Shrestha et al., 2012*)	Ontology	Plant	Validated trait names, annotation of phenotypic and genotypic data	Eight specific crop ontologies and trait dictionaries (cassava, chickpea, common bean, groundnut, maize, Musa, potato, rice, sorghum, and wheat).	http://www.cropontology.org/
Planteome (*Cooper et al., 2018*)	Ontology	Plant	Reference ontologies, plant genomics and phenomics	95 plant taxa	http://www.planteome.org
Solanum tuberosum knowledge base (*Ivanisenko et al., 2018*)	Literature mining	Potato, Maize, Rice, Arabidopsis thaliana	Molecular genetic regulation and metabolic pathways	9,000 full-text articles and more than 130,000 abstracts from PubMed	http://www-bionet.sysbio.cytogen.ru/and/plant/
(*Singh et al., 2021*)	Literature mining	Potato, Tomato, Eggplant, Capsicum	Flesh color phenotypic trait knowledge network	4,023 PubMed abstracts of plant genetics-based articles	
CIP (*Mihovilovich et al., 2015*)		Potato	Germplasm bank and Gene Resource Bank		https://cipotato.org/genebankcip/
AHDB		Potato	Various aspects of potato production, including seed health, agronomy, crop protection, storage, and crop quality		https://potatoes.ahdb.org.uk/

(CIP; *Mihovilovich et al. (2015)*) and the UK Agriculture and Horticulture Development Board (AHDB) also provide abundant potato germplasm resources. As shown in Table 1, existing knowledge bases constructed through literature mining do not include a potato gene-disease knowledge base.

To provide the associations between potato genes and diseases, we used natural language processing (NLP) technology and large language models (LLM) to construct the Potato Gene-Disease Knowledge Base.

There are three primary methods of extracting knowledge from scientific literature: manual extraction, rule-based extraction, and machine learning. The manual method, although highly accurate, is extremely time-consuming (*Xing et al., 2018*). Rule-based and template-based methods heavily rely on biological experts, as illustrated by *Choi et al. (2016)*, *Larmande, Do & Wang (2019)*, *Saik et al. (2017)*. The traditional machine learning approach heavily relies on manually annotated training data (*Mausam, 2016*; *Fiorini et al., 2018*). Methods based on LLM are increasingly being used and have achieved state-of-the-art results in various fields (*Wadhwa, Amir & Wallace, 2023*; *Sun et al., 2024*).

This article introduces the PotatoG-DKB, which was constructed using literature mining and natural language processing technology to extract biological entities such as genes, diseases, fungi, bacteria, and pests. LLM was used to determine the semantic relationships between these entities. These relationships were then mapped onto a Potato Gene-Disease Knowledge Graph (PotatoG-DKG). This multi-relational graph includes nodes that represent biological entities and edges that signify relationships between these entities. The PotatoG-DKG serves as a comprehensive knowledge base that integrates diverse information and displays correlations in a graph structure. This format simplifies access to relevant knowledge and can support research efforts. We also developed a visualization platform for PotatoG-DKB using web technology. The Potato Gene-Disease Knowledge Portal (PotatoG-DKP) allows biologists and potato breeders to intuitively explore information embedded in the literature.

MATERIALS AND METHODS

In this study, scientific literature served as the primary source of experimental data. To construct the PotatoG-DKB, unstructured information was extracted from potato-related scientific literature using a range of techniques. First, the query "potato and gene and disease" was used on PubMed (https://pubmed.ncbi.nlm.nih.gov/) to acquire 2,906 literature abstracts. The PubTator tool was then used to extract the basic biological entities, such as genes, proteins, species, and mutations, from these abstracts. Other biological entities were extracted using a pattern-matching approach. To improve efficiency, the text of the original abstracts was broken into sentences, and sentences containing at least two biological entities were retained. A large language model (LLM) with constructed prompts was used to extract data on potato gene-disease relationships and other related information. To ensure data accuracy, the obtained data was manually curated into PotatoG-DKB. The structured data was stored in a Neo4j graph database. To enable potato researchers and breeders to quickly access relevant knowledge, we developed the PotatoG-DKP. The construction process of the PotatoG-DKP is illustrated in Fig. 1.

Preprocessing

The data sources analyzed in this article were retrieved from PubMed (https://pubmed.ncbi.nlm.nih.gov/), a search engine maintained by the National Center for Biotechnology

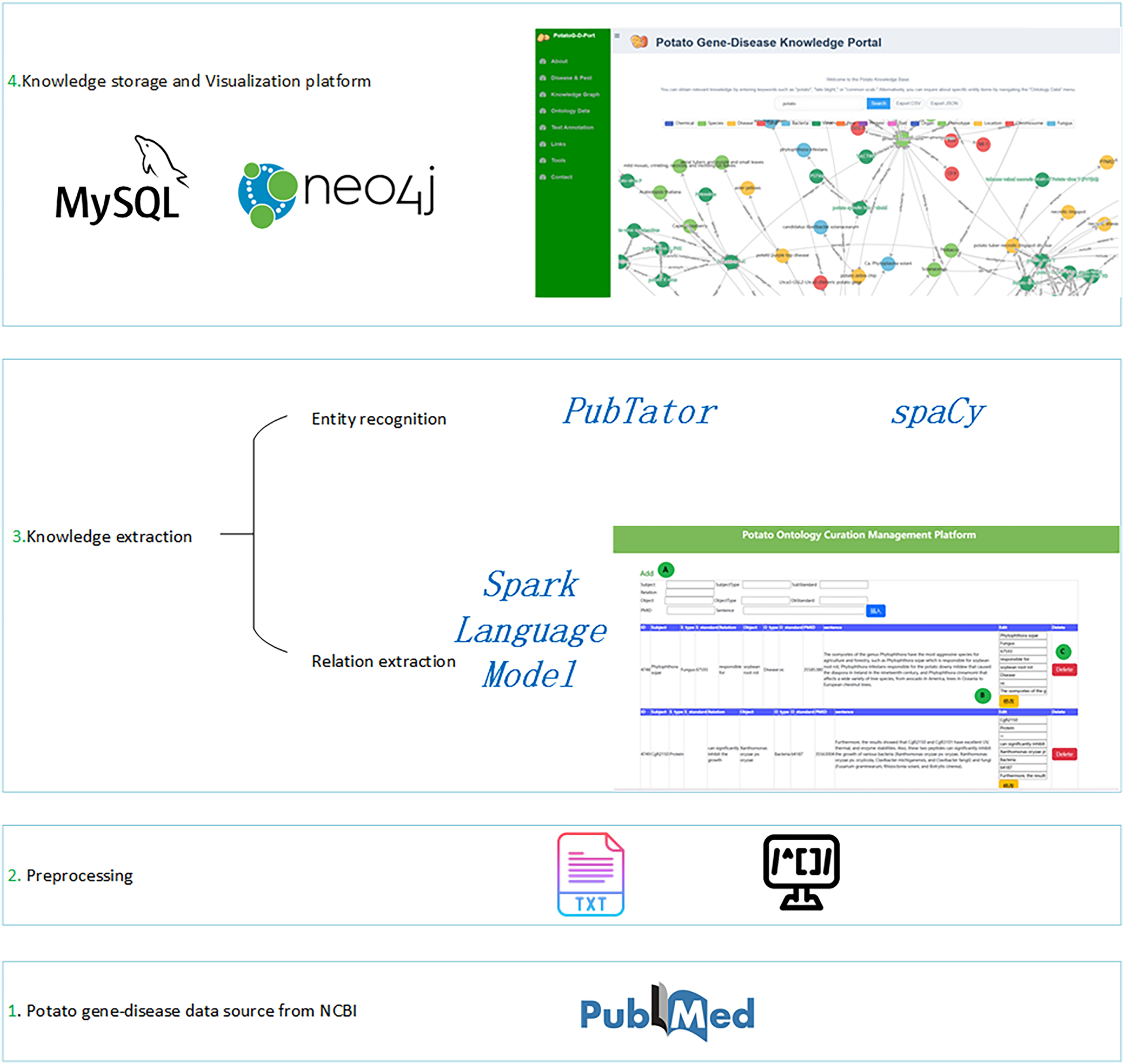

4.Knowledge storage and Visualization platform
MySQL
neo4j
Potato Gene-Disease Knowledge Portal
3.Knowledge extraction
Entity recognition
Relation extraction
PubTator
spaCy
Spark Language Model
Potato Ontology Curation Management Platform
2. Preprocessing
TXT
1. Potato gene-disease data source from NCBI
PubMed

Figure 1 The flowchart for constructing the potato gene-disease knowledge base. Txt icon designed by iconixar on Flaticon.com. Coding icon designed by orvipixel on Flaticon.com. Full-size DOI: 10.7717/peerj.18202/fig-1

Information (NCBI) at the National Library of Medicine (NLM; *Wei et al. (2019)*). For this study, "potato and gene and disease" was used as the search string in PubMed, and 2,906 final abstracts were selected from the search results. All sources were saved in the PubMed_ID (PMID) format. PubTator Central provides automatic annotations of biological concepts from PubMed abstracts. The PubTator API (https://www.ncbi.nlm.nih.gov/research/pubtator/api.html) provides source code downloads in Perl, Python, and

Java, and the annotation formats include PubTator, biocxml, and biocjson. PubTator was used to annotate the obtained sources in PubTator format, including titles, abstracts, and annotations of biological entities, saved in ".txt" format. We used rule-based syntax in sentence segmentation to avoid inaccuracies associated with punctuation-based splitting (https://docs.python.org/3/library/re.html). For example, rule-based syntax prevents the incorrect segmentation of "*B. lactucae* regulatory sequences in *P. infestans*" into "B. | lactucae regulatory sequences in P. | infestans." A dictionary of common potato pests and diseases was then used with SpaCy's pattern-matching functionality to expand the entity terms to enhance named entity recognition (https://spacy.io/). All sentences not containing entity information were removed, leaving a total of 21,995 sentences containing at least biological two entities.

Named entity recognition

Knowledge extraction, a crucial aspect of knowledge base construction, encompasses both named entity recognition (NER) and relationship extraction (RE). NER in the field of biology primarily targets genes, diseases, chemicals, species, and mutations. PubTator Central (PTC; *Wei et al. (2019)*) was used in this study for biological entity recognition (*e.g*., genes, diseases, chemicals, species, mutations, and cell lines). PTC, a web-based system, offers a multi-entity recognition interface—PubTator Restful API (https://www.ncbi.nlm.nih.gov/research/pubtator/api.html)—to export annotations. The named entity recognition annotations are displayed in Table 2. PMID denotes PubMed's unique identifier, "start" and "end" indicate the boundaries of mentions in the abstract, "mention" refers to the biological entity, "category" specifies the type of biological entity, and "standard" refers to standardized identification information of the biological entities.

In PubTator, diseases are identified based on Disease Ontology (*Schriml et al., 2012*), which is a database of human diseases. Because of this, PubTator's accuracy in recognizing crop-related diseases is relatively low. To address this issue, we enhanced disease and pest entity recognition using dictionary-based and pattern-matching methods, implemented through SpaCy's entity recognition tool (https://spacy.io/). SpaCy is a free, open-source library for advanced NLP in Python.

PubTator classifies pathogen information for potatoes under "species." To provide a more detailed representation of potato diseases and pests, we further categorized potato pathogens into fungal pathogens, bacterial pathogens, viral pathogens, and pests. Additional entities, such as abiotic stress and chromosome, were incorporated as needed to enrich the relationships within PotatoG-DKB. Ultimately, we identified data for 22 categories of entities. The detailed information on these categories is shown in Table 3.

Entities are represented in various forms in the literature, including abbreviations, word variants, and synonyms. For instance, "*Phytophthora infestans*" can also be referred to as "*P. infestans*", "*Phytophothora*", "*Pi*", or "*Phytophthora infestans* (Mont.) de Bary". Therefore, unique identifiers had to be established for each entity. Unique identifiers for entities such as genes, proteins, chemicals, and mutations were sourced from PubTator, while identifiers for species, fungi, bacteria, viruses, and pests were obtained from the Taxonomy database (https://www.ncbi.nlm.nih.gov/taxonomy). Entities without unique

Table 2 Examples of entity annotation using PubTator.

PMID	Start	End	Mention	Category	Standard
1351246	51	73	*Phytophthora infestans*	Species	4787
1351246	77	83	Potato	Species	4113
25681825	623	649	Potato late blight disease	Disease	MESH:D055750
2132026	975	979	Prp1	Gene	100795066
1655113	860	870	Asparagine	Chemical	MESH:D001216
25873665	1050	1055	V314I	Mutation	tmVar:p|SUB|V|314|I;HGVS:p.V314I;VariantGroup:1

Table 3 Brief definition and examples for each of the 22 categories in the entity annotation.

S.No.	Annotation category	Brief definition	Examples
1	Disease	Diseases caused by potato infection pathogens	Late blight, soft rot
2	Fungus	Potato fungi pathogenic	*Phytophthora infestans*
3	Bacteria	Potato bacterial pathogens	*Pseudomonas syringae*
4	Virus	Potato virus pathogen	Potato virus Y, tomato mosaic virus
5	Pest	Potato pest pathogen	Potato cyst nematode
6	Species	Categorize of taxonomy database	Potato, tomato
7	Gene	Name of genetic factor	STWRKY8, 3-hydroxy-3-methylglutaryl-coenzyme a reductase
8	Mutation	Changes in nucleotide sequence of organism, virus or chromosome DNA genome	PiGPA1-deficient mutants
9	Chemical	Compounds involved in molecular biology research	Salicylic acid
10	Abiotic stress	The adverse effects of non-biological factors in the environment on potato plants	Drought, cold, heat
11	Biotic stress	The general term for biological stress.	Biotic stress
12	Protein	Protein, elicitin, enzyme	PR1, mitogen-activated protein kinase
13	Chromosome	Potato chromosome	Chromosome 8
14	Organ	Organ of plant	Leaf, stem, tuber
15	Tissue	Type of cell or part of cell	Cytoplasm
16	Phenotype	The observable traits or characteristics of an organism	Violet flower colour
17	Structure	The structure of proteins	1,134 amino acid residues
18	Trait	Measurable features	Early maturity, late blight resistance
19	mRNA	Messenger RNA	Star mRNA
20	Location	Pest and disease occurrence location	Mexico
21	QTL	Quantitative trait locus	GpaVvrn
22	Molecular biology tool	Equipment or software used in molecular biology experiments and research	RSOLA

identifiers are represented as '–'. A potato-related entity dictionary was then constructed based on these entities and their categories. Examples from the entity dictionary are shown in Table 4, where "mention" indicates a meaningful item identified in potato-related

Table 4 **The examples of entity dictionary.**

Mention	Category	Unique identifier
Phytophthora infestans	Species	4787
Potato	Species	4113
Potato late blight	Disease	–
Prp1	Gene	100795066
Asparagine	Chemical	MESH:D001216
Erwinia carotovora	Bacteria	554
V314I	Mutation	tmVar:p|SUB|V|314|I;HGVS:p.V314I;VariantGroup:1
Potato psyllid	Pest	290155
Alternaria solani	Fungus	48100
Tomato leaf curl New Delhi virus	Virus	223347
Tuber	Organ	–
Low temperature stress	Abiotic stress	–

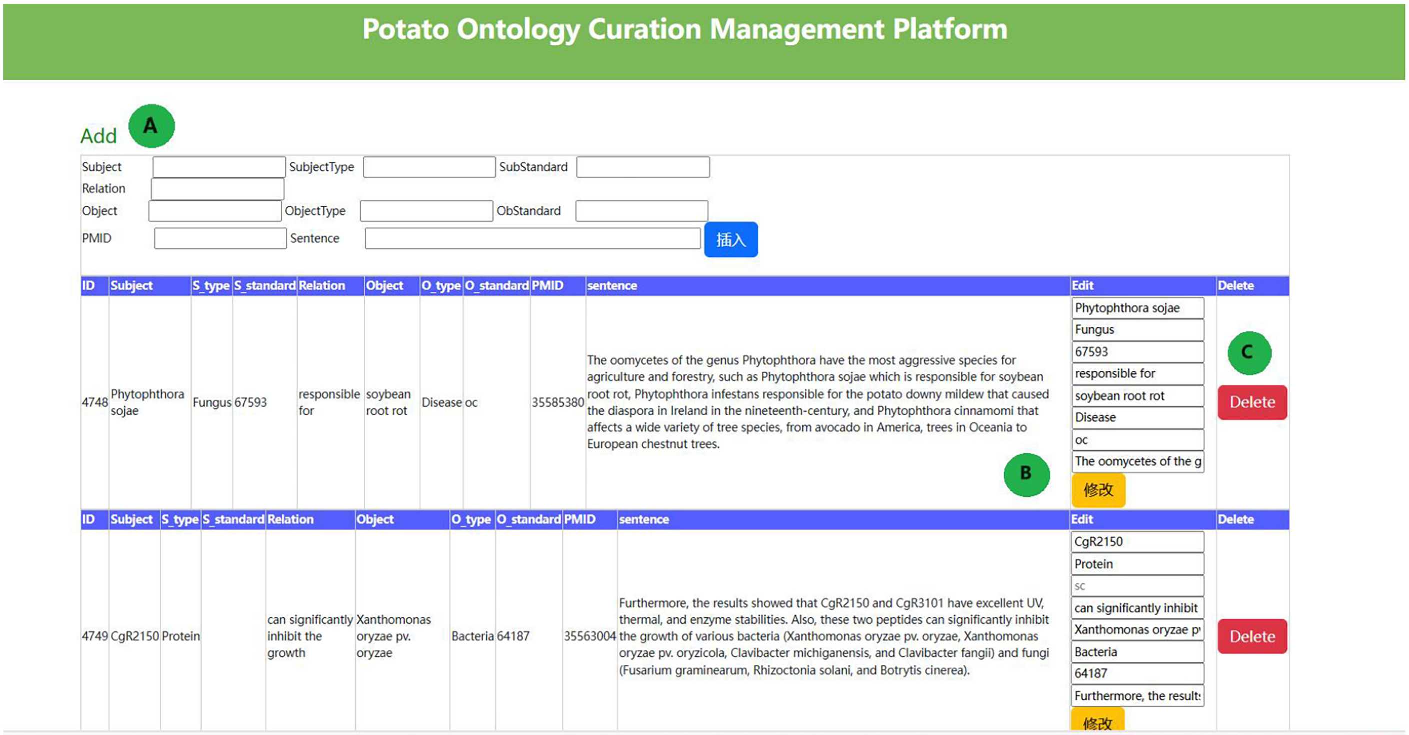
Table 5 **LLM for relation extraction.** During the named entity recognition stage, we incorporate the identified entities into the prompt template as input for the LLM, enabling us to retrieve corresponding relationship information.

Original sentence	Recognized entity information
Late blight in potato is caused by the fungus *Phytophthora infestans* and can inflict severe damage on the potato crop	[‘late blight’, ‘potato’, ‘*Phytophthora infestans*’]
Prompt	
In the sentence “Late blight in potato is caused by the fungus *Phytophthora infestans* and can inflict severe damage on the potato crop ”, we have obtained the entities of [‘late blight’, ‘potato’, ‘*Phytophthora infestans*’]. Please provide the relationship formed by these entities in the format of {subject, relationship, object}	
LLM output	
{“Late blight”, “is a disease of”, “potato”} {“*Phytophthora infestans*”, “causes”, ,“late blight”} {“*Phytophthora infestans*”, “can inflict damage on”, “potato crop”}	

literature, “category” refers to the type of entity, and “unique identifier” is the unique identifier of the entity in the NCBI database.

Relation extraction

Relation extraction (RE) is a subdomain of syntactic analysis aimed at revealing the semantic relationships between entities in unstructured text. In natural language processing, RE is typically treated as a supervised classification task. To achieve high accuracy in relationship extraction, a large amount of training data is required. Unfortunately, there is currently no specialized biological text *corpus* specific to potatoes. With the introduction of ChatGPT (OpenAI, 2022), large language models based on generative frameworks have shown promising performance across various tasks. In this study, LLM was used to extract triplet relationships. To enhance the efficiency of relationship extraction, we retained sentences containing at least two recognized entities.



Potato Ontology Curation Management Platform

Subject | SubjectType | SubStandard
Relation
Object | ObjectType | ObStandard
PMID | Sentence

ID	Subject	S_type	S_standard	Relation	Object	O_type	O_standard	PMID	sentence	Edit	Delete
4748	Phytophthora sojae	Fungus	67593	responsible for	soybean root rot	Disease	oc	35585380	The oomycetes of the genus Phytophthora have the most aggressive species for agriculture and forestry, such as Phytophthora sojae which is responsible for soybean root rot, Phytophthora infestans responsible for the potato downy mildew that caused the diaspora in Ireland in the nineteenth-century, and Phytophthora cinnamomi that affects a wide variety of tree species, from avocado in America, trees in Oceania to European chestnut trees.	Phytophthora sojae Fungus 67593 responsible for soybean root rot Disease oc The oomycetes of the g 修改	Delete

ID	Subject	S_type	S_standard	Relation	Object	O_type	O_standard	PMID	sentence	Edit	Delete
4749	CgR2150	Protein		can significantly inhibit the growth	Xanthomonas oryzae pv. oryzae	Bacteria	64187	35563004	Furthermore, the results showed that CgR2150 and CgR3101 have excellent UV, thermal, and enzyme stabilities. Also, these two peptides can significantly inhibit the growth of various bacteria (Xanthomonas oryzae pv. oryzae, Xanthomonas oryzae pv. oryzicola, Clavibacter michiganensis, and Clavibacter fangii) and fungi (Fusarium graminearum, Rhizoctonia solani, and Botrytis cinerea).	CgR2150 Protein sc can significantly inhibit Xanthomonas oryzae p Bacteria 64187 Furthermore, the result: 修改	Delete

Figure 2 **The potato ontology curation management platform.** Full-size DOI: 10.7717/peerj.18202/fig-2

By constructing feature templates, we extracted the relevant relationships between entities using the Spark Language Model V3.5, a large language model released by iFlytek (https://xinghuo.xfyun.cn/sparkapi). An example of relationship extraction using this large language model is shown in Table 5. The extracted relationships, PMID, and sentences were then stored in a MySQL database.

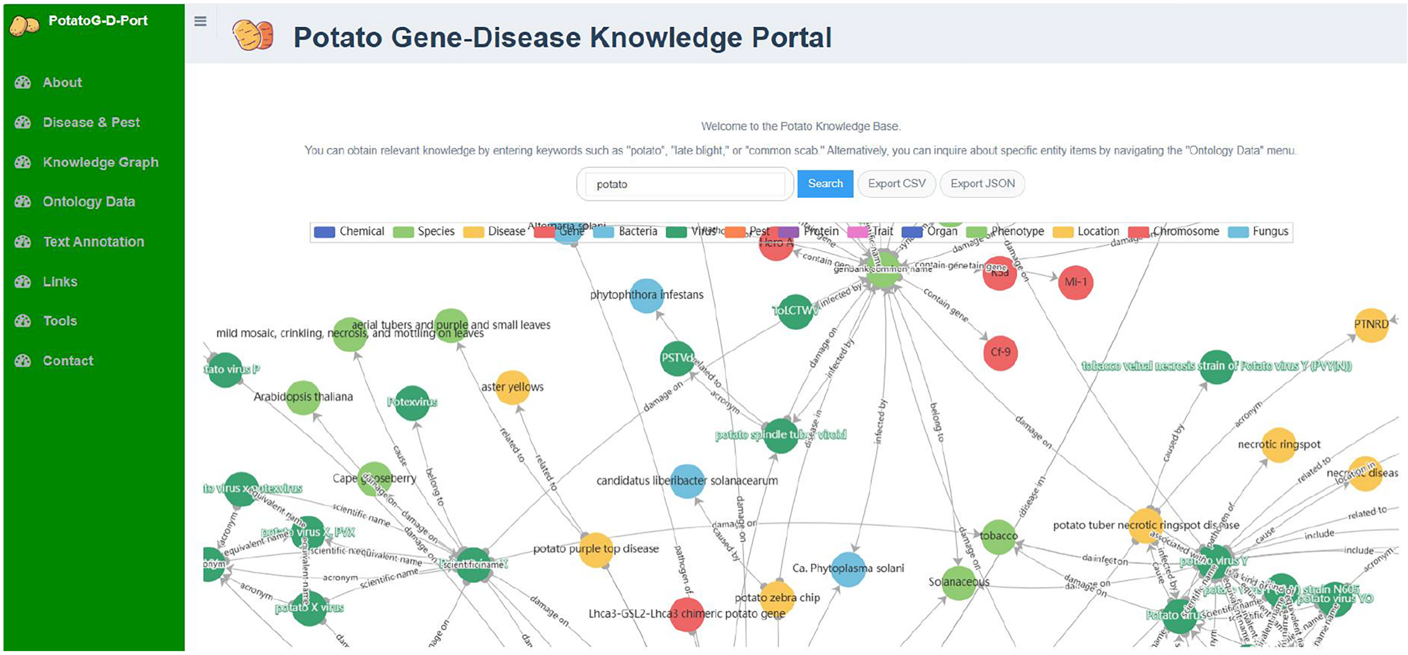
We constructed a potato ontology by combining entity dictionaries and relational data. It is widely acknowledged that the co-occurrence of biological entities in a sentence indicates a strong correlation. LLM can accurately extract triplet information for highly correlated entities. However, their performance is poorer for entities with weaker correlations, resulting in incomplete relationship extraction. We developed a potato ontology platform to correct erroneous results, remove irrelevant data, and manually add relationships for those unrecognized. The potato ontology curation platform is illustrated in Fig. 2. This platform resulted in a more refined potato knowledge base, which includes 5,701 relational tuples. The species, fungi, bacteria, viruses, and pests were also expanded in the potato knowledge base using taxonomy dictionaries (https://www.ncbi.nlm.nih.gov/taxonomy), resulting in the addition of 5,003 new relationships and final total of 10,704 relationship data points. The relationship data between entities is shown in Table 6.

Knowledge storage and visualization platform

Neo4j is an open-source graph database that stores structured knowledge in the form of graphs rather than traditional tables and supports semantic queries

Table 6 The examples of relationship between entities. In the relationship table, the data in the first section is obtained through LLM, while the data in the second section is expanded using the Taxonomy dictionary.

Mention1	Relation	Mention2
Prp1	Resistance to	*Phytophthora infestans*
Ralstonia solanacearum	Cause	Wilt
Snakin-1	Gene in	Potato
Early blight QTL	On	Chromosome 2
Late blight	Caused by	*Phytophthora infestans*
Stenotrophomonas maltophilia	Resistance to	*Ralstonia solanacearum*
Candidatus Liberibacter psyllaurous	Cause	Zebra chip
Quinone outside inhibitor (QoI) fungicides	Control	Brown spot
Gibberella pulicaris	Synonym	Fusarium roseum
Gibberella pulicaris	Authority	Fusarium roseum link, 1809
Gibberella pulicaris	Scientific name	Fusarium sambucinum
Gibberella pulicaris	Authority	*Gibberella pulicaris* (Kunze) Sacc., 1877



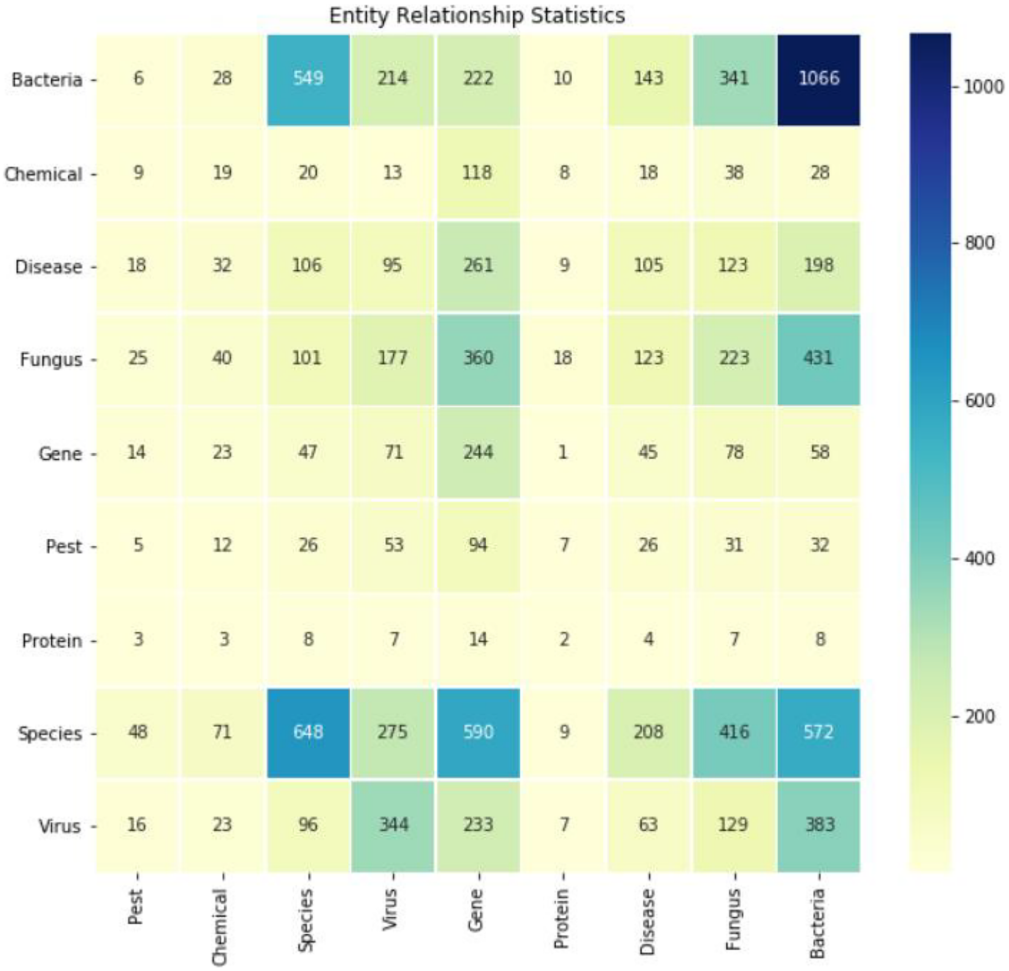
Welcome to the Potato Knowledge Base.

You can obtain relevant knowledge by entering keywords such as "potato", "late blight," or "common scab." Alternatively, you can inquire about specific entity items by navigating the "Ontology Data" menu.

potato

Figure 3 The potato gene-disease knowledge graph visualization platform. Full-size DOI: 10.7717/peerj.18202/fig-3

(*Holzschuher & Peinl, 2013*). We chose Neo4j as the graph database for PotatoG-DKB. The Potato Gene-Disease Knowledge Portal (PotatoG-DKP) was developed using Django as the front-end interface, a high-level Python web framework, with Neo4j as the graph database and a Neo4j + MySQL combination as the back-end implementation. The PotatoG-DKP supports visual query, as shown in Fig. 3. Users can input an entity of



	Pest	Chemical	Species	Virus	Gene	Protein	Disease	Fungus	Bacteria
Bacteria	6	28	549	214	222	10	143	341	1066
Chemical	9	19	20	13	118	8	18	38	28
Disease	18	32	106	95	261	9	105	123	198
Fungus	25	40	101	177	360	18	123	223	431
Gene	14	23	47	71	244	1	45	78	58
Pest	5	12	26	53	94	7	26	31	32
Protein	3	3	8	7	14	2	4	7	8
Species	48	71	648	275	590	9	208	416	572
Virus	16	23	96	344	233	7	63	129	383

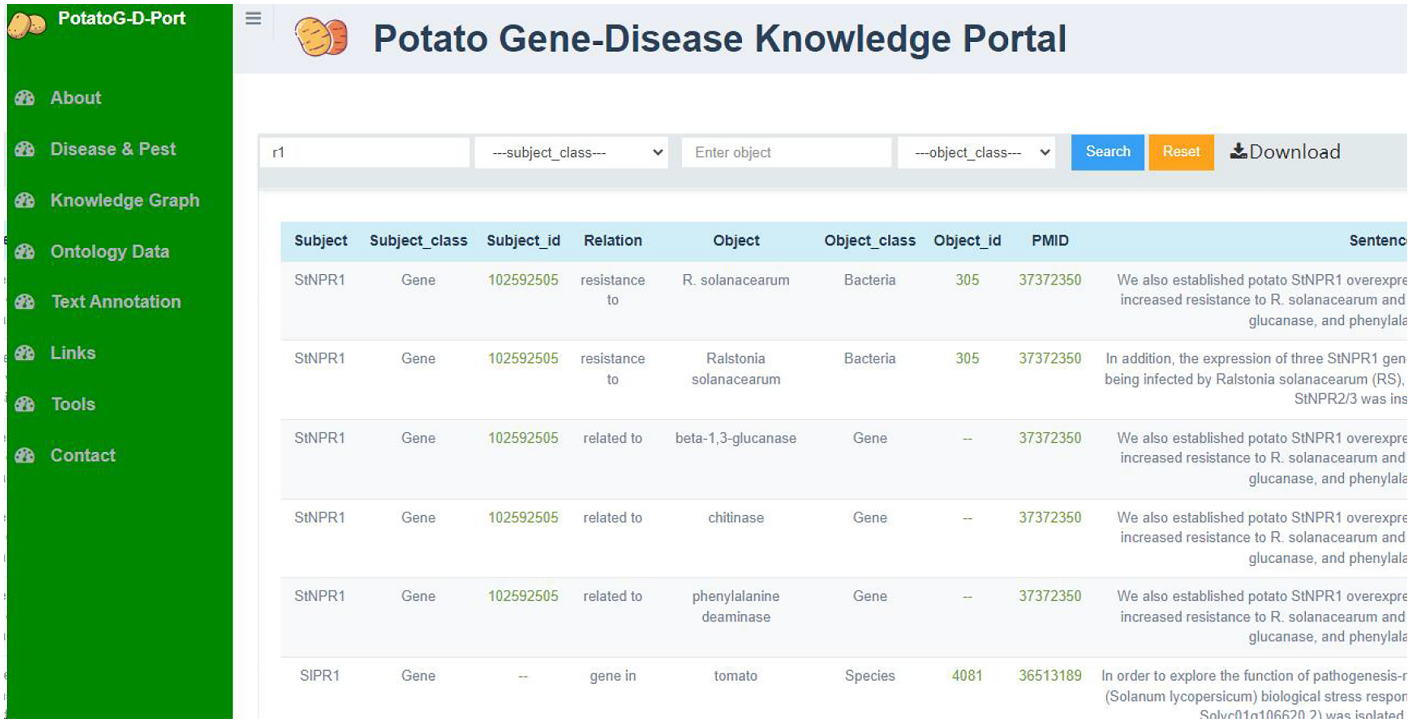
Figure 4 **Heatmap of several typical relationships extracted (*e.g.*, gene-disease, gene-gene, bacteria-bacteria).** Full-size DOI: 10.7717/peerj.18202/fig-4

interest to query related entities and relationships. A recursive method is used to support querying extended nodes and queried data can be exported in CSV (Comma-Separated Values) and JSON (JavaScript Object Notation) formats. If users do not input any query information, the portal can display the entirety of PotatoG-DKG. To ensure data security, a full data download option is not provided, but researchers can contact us for more information if they need access to the complete data.

RESULTS

Named entity recognition and relationship extraction

PubTator was used for named entity recognition and the Spacy tool was used for entity identification. Using the taxonomy dictionary to expand entities, we obtained a total of 5,206 entity items, including 300 genes, 205 diseases, 300 fungi, 300 bacteria, 109 pests, 300 viruses, 300 species, 277 chemicals, and 73 proteins as well as other entities.



Subject	Subject_class	Subject_id	Relation	Object	Object_class	Object_id	PMID	Sentenc
StNPR1	Gene	102592505	resistance to	R. solanacearum	Bacteria	305	37372350	We also established potato StNPR1 overexpre increased resistance to R. solanacearum and glucanase, and phenylala
StNPR1	Gene	102592505	resistance to	Ralstonia solanacearum	Bacteria	305	37372350	In addition, the expression of three StNPR1 gen being infected by Ralstonia solanacearum (RS), StNPR2/3 was ins
StNPR1	Gene	102592505	related to	beta-1,3-glucanase	Gene	--	37372350	We also established potato StNPR1 overexpre increased resistance to R. solanacearum and glucanase, and phenylala
StNPR1	Gene	102592505	related to	chitinase	Gene	--	37372350	We also established potato StNPR1 overexpre increased resistance to R. solanacearum and glucanase, and phenylala
StNPR1	Gene	102592505	related to	phenylalanine deaminase	Gene	--	37372350	We also established potato StNPR1 overexpre increased resistance to R. solanacearum and glucanase, and phenylala
SlPR1	Gene	--	gene in	tomato	Species	4081	36513189	In order to explore the function of pathogenesis-r (Solanum lycopersicum) biological stress respon Solyc01g106620.2) was isolated

Figure 5 **The ontology information of potato gene-disease knowledge base.** Full-size DOI: 10.7717/peerj.18202/fig-5

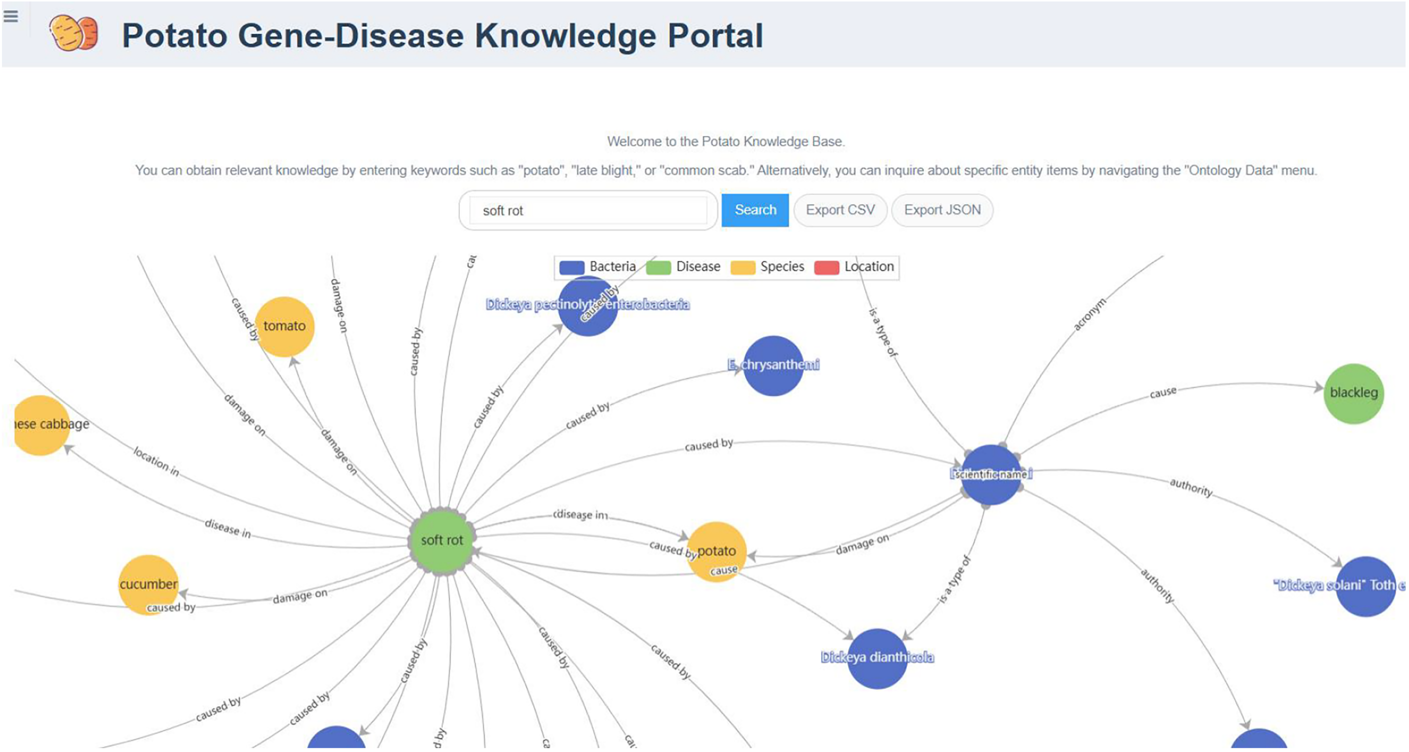
The Spark Large Language Model was used for relationship extraction, supplemented by manual curation. This process yielded a total of 10,704 relationship tuples. Figure 4 depicts the heatmap of typical relationships such as "gene-disease" and "fungus-disease." Each row and column respectively represent a category of interest (*e.g.*, genes, diseases, species), and each cell in the matrix represents the strength or frequency of the relationship between the two categories. The darker the color, the stronger the correlation between the two entities.

Potato gene-disease knowledge base

Through the potato ontology data curation platform, we obtained potato ontology data on gene-disease relationships in potato as well as other relationships between biological entities, which, together, constitute the PotatoG-DKB, as shown in Fig. 5. This knowledge base includes entities, entity_categories, entity_id, relationships, PMIDs, and sentence information describing relevant relationships. Users can browse, search, and download relevant data. Additionally, users can retrieve information based on entity items or entity categories according to their needs. The PotatoG-DKB also provides an information tracing function, allowing users to link to the original PubMed literature through PMIDs or query detailed information about entities in NCBI using entity IDs. More detailed information can be accessed and downloaded from https://www.potatogd.com.cn.

Potato gene-disease knowledge graph

The PotatoG-DKG covers 5,206 nodes and 9,443 edges. To make scientific literature directly and readily accessible to biologists and potato breeders, we developed a



Potato Gene-Disease Knowledge Portal

Welcome to the Potato Knowledge Base.

You can obtain relevant knowledge by entering keywords such as "potato", "late blight," or "common scab." Alternatively, you can inquire about specific entity items by navigating the "Ontology Data" menu.

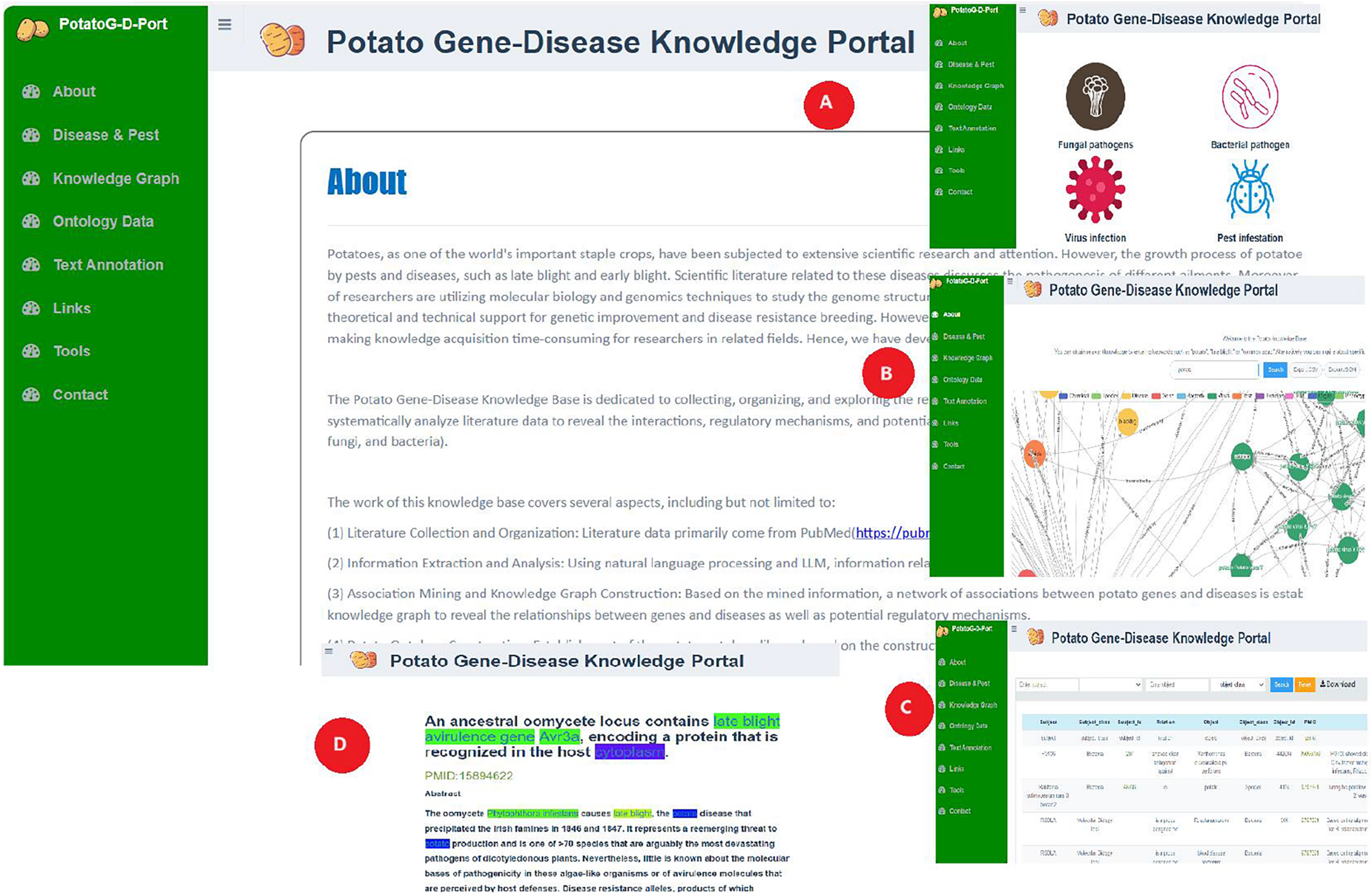
soft rot | Search | Export CSV | Export JSON

Figure 6 **The relationship graph returned by searching with the keyword "soft rot".** Full-size DOI: 10.7717/peerj.18202/fig-6

visualization platform, depicted in Fig. 3. The search function of the Neo4j graph database directly returns relevant relationships. For example, when searching for "A", it may return "A->B". In order to expand the retrieved relationships, a recursive search algorithm was adopted with a recursion depth limit of five to prevent excessive search depth and subsequent inefficiencies in querying. Figure 6 demonstrates the search results returned when using the keyword "soft rot" to search for relevant knowledge. A data download function is also provided, allowing users to export the data in both CSV and JSON formats.

Potato gene-disease knowledge portal

Relevant researchers and breeders can access and query potato gene-disease ontology data through the PotatoG-DKP. As shown in Fig. 7, the "about" section of the portal website introduces the main process of potato gene-disease knowledge base construction. The "disease and pest" section displays common potato diseases caused by fungi, bacteria, viruses, and pests (Fig. 7A). The "knowledge graph" module enables visual querying of relevant information in the ontology database. To facilitate user operations, PotatoG-DKP has a fuzzy search function, and users can also download the entity-relationship information of interest (Fig. 7B). In the "ontology" section, PotatoG-DKP provides potato gene-disease ontology data, where users can also query or download relevant data (Fig. 7C). The "text annotation" module offers text annotation, allowing users to search and read abstracts of relevant literature (Fig. 7D). If they wish to delve deeper into the



About

Potatoes, as one of the world's important staple crops, have been subjected to extensive scientific research and attention. However, the growth process of potatoe
by pests and diseases, such as late blight and early blight. Scientific literature related to these diseases
of researchers are utilizing molecular biology and genomics techniques to study the genome structur
theoretical and technical support for genetic improvement and disease resistance breeding. Howeve
making knowledge acquisition time-consuming for researchers in related fields. Hence, we have deve

The Potato Gene-Disease Knowledge Base is dedicated to collecting, organizing, and exploring the re
systematically analyze literature data to reveal the interactions, regulatory mechanisms, and potentia
fungi, and bacteria).

The work of this knowledge base covers several aspects, including but not limited to:

(1) Literature Collection and Organization: Literature data primarily come from PubMed(https://pubr

(2) Information Extraction and Analysis: Using natural language processing and LLM, information rela

(3) Association Mining and Knowledge Graph Construction: Based on the mined information, a network of associations between potato genes and diseases is estab
knowledge graph to reveal the relationships between genes and diseases as well as potential regulatory mechanisms.

on the construc

An ancestral oomycete locus contains late blight avirulence gene Avr3a, encoding a protein that is recognized in the host cytoplasm.

PMID:15894622

Abstract

The oomycete Phytophthora infestans causes late blight, the potato disease that precipitated the Irish famines in 1846 and 1847. It represents a reemerging threat to potato production and is one of >70 species that are arguably the most devastating pathogens of dicotyledonous plants. Nevertheless, little is known about the molecular bases of pathogenicity in these algae-like organisms or of avirulence molecules that are perceived by host defenses. Disease resistance alleles, products of which

Figure 7 **Potato gene-disease knowledge portal schematic diagram.** Full-size DOI: 10.7717/peerj.18202/fig-7

specific content, they can directly access the original PubMed article through the PMID. "Links" to related databases are also provided, and users can access or download the tools used in this article through the "tools" section.

DISCUSSION

The Potato Gene-Disease Knowledge Graph visualizes the network of related entities in a graphical manner, aiding in the discovery of hidden relationships between knowledge elements. Figure 8 illustrates the results obtained from searching for the gene entity named "Rx2" within the knowledge graph. These results include connections such as {"Rx2", "is similar to", "Rx1"}, {"Rx2", "is similar to", "Rxh1"}, {"Rxh1", "encodes", "Gpa2"}, and {"Gpa2", "resistance to", "*Globodera rostochiensis*"}. Although "Rx1" is not directly linked to "*Globodera rostochiensis*", "Rx1" may also confer resistance to this pathogen (as indicated by the green dashed lines in the figure); the relationship between "Rx2" and "*Globodera rostochiensis*" warrants further investigation. More information is available at https://www.potatogd.com.cn.

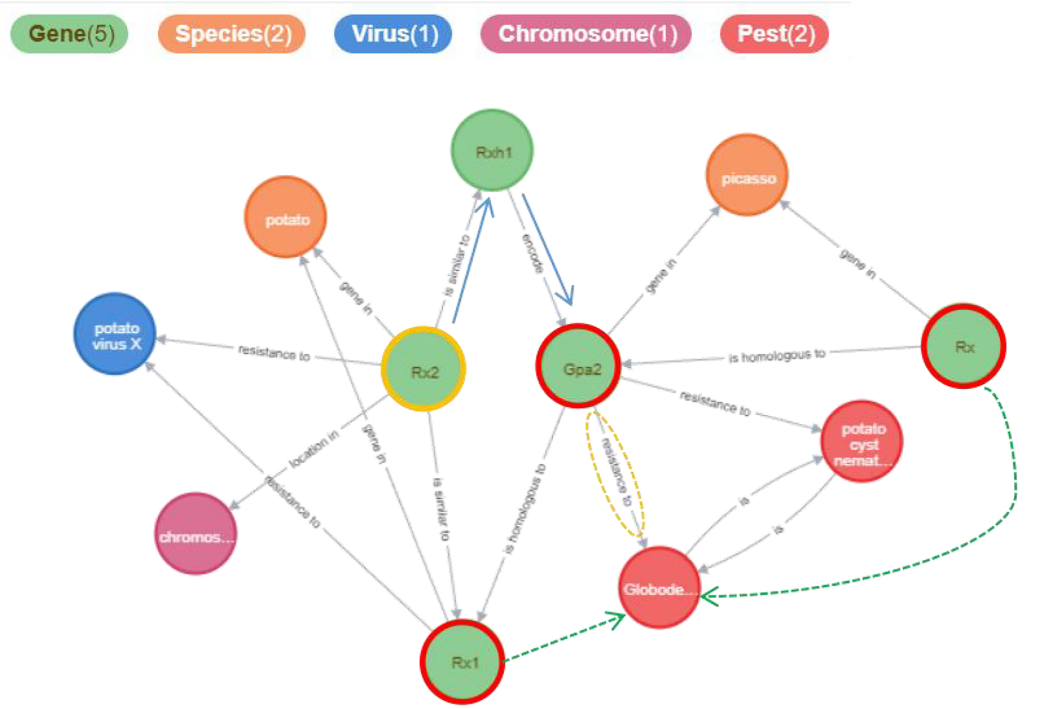


Figure 8 **The results returned from searching for the "Rx2" relationship using keywords in PotatoG-DKG.** We presented the relevant relationships of genes in the figure, with the green dashed lines highlighting relationships that are worthy of further investigation by researchers.

Full-size DOI: 10.7717/peerj.18202/fig-8

CONCLUSIONS

This study used natural language processing technology and LLMs to analyze literature pertaining to potato gene-disease relationships. By performing entity recognition and relationship extraction, we constructed a knowledge graph comprising 5,206 nodes and 9,443 edges. To facilitate researchers in the use of this information, we developed a potato gene-disease knowledge portal (PotatoG-DKP) capable of effectively organizing and visually presenting the associated relationships. The findings of this study suggest that leveraging associations extracted from scientific literature can assist in uncovering hidden relationships among entities like genes and diseases, and integrating such data may foster the discovery of new hypotheses.

We will continuously update the Potato Gene-Disease Knowledge Base as new research articles are published, ensuring the data remains current and comprehensive. The raw data we used were article abstracts from PubMed, which reflect the key information of the article, but do not include the entirety of the information contained in the article. Therefore, in the future, we will consider processing the PubMed Central full-text database. By integrating multi-source data, we aim to enrich the knowledge base. Additionally, we plan to develop an intelligent question-answering assistant for potato gene-disease relationships based on the constructed PotatoG-DKB and LLM, further aiding potato researchers.

ACKNOWLEDGEMENTS

We would like to thank Professor Zhongren Yang, a potato expert from the Inner Mongolia Agricultural University and Vice President of National Innovation Alliance of Northeast Mountain Wild Vegetable Industry, and his team for their help and guidance in data collation and review. We would also like to thank Professor Wenguang Zhang, a bioinformatics specialist of the College of Life Science, Inner Mongolia Agricultural University, for his guidance in data review.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was supported by the Inner Mongolia Science and Technology Major Special Projects [2021ZD0005]; Natural Science Foundation of Inner Mongolia Autonomous Region [2020MS06013]; Inner Mongolia Agricultural University [JC2019005]; Basic Scientific Research Business Project of Universities in Inner Mongolia Autonomous Region [BR221022]; Inner Mongolia Agricultural University College Students' Innovation and Entrepreneurship Training Program [202210129003]. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
Inner Mongolia Science and Technology Major Special Projects: 2021ZD0005.
Natural Science Foundation of Inner Mongolia Autonomous Region: 2020MS06013.
Inner Mongolia Agricultural University: JC2019005.
Basic Scientific Research Business Project of Universities in Inner Mongolia Autonomous Region: BR221022.
Inner Mongolia Agricultural University College Students' Innovation and Entrepreneurship Training Program: 202210129003.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Congjiao Xie conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the article, and approved the final draft.
- Jing Gao conceived and designed the experiments, analyzed the data, authored or reviewed drafts of the article, and approved the final draft.
- Junjie Chen analyzed the data, authored or reviewed drafts of the article, and approved the final draft.
- Xuyang Zhao analyzed the data, prepared figures and/or tables, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

The data is available at the Potato Gene-Disease Knowledge Portal: https://potatogd.com.cn; and at Zenodo: Xie, C. (2024). potato gene-disease knowledge base [Data set]. Zenodo. https://doi.org/10.5281/zenodo.12706091.

REFERENCES

Bolser DM, Staines DM, Perry E, Kersey PJ. 2017. Ensembl plants: integrating tools for visualizing, mining, and analyzing plant genomic data. *Plant Genomics Databases: Methods and Protocols* **1533**:1–31 DOI 10.1007/978-1-4939-6658-5.

Choe J, Kim JE, Lee BW, Lee JH, Nam M, Park YI, Jo SH. 2018. A comparative synteny analysis tool for target-gene SNP marker discovery: connecting genomics data to breeding in Solanaceae. *Database* **2018**:bay047 DOI 10.1093/database/bay047.

Choi W, Kim B, Cho H, Lee D, Lee H. 2016. A corpus for plant-chemical relationships in the biomedical domain. *BMC Bioinformatics* **17(1)**:1–15 DOI 10.1186/s12859-016-1249-5.

Cooper L, Meier A, Laporte MA, Elser JL, Mungall C, Sinn BT, Cavaliere D, Carbon S, Dunn NA, Smith B, Qu B, Preece J, Zhang E, Todorovic S, Gkoutos G, Doonan JH, Stevenson DW, Arnaud E, Jaiswal P. 2018. The Planteome database: an integrated resource for reference ontologies, plant genomics and phenomics. *Nucleic Acids Research* **46(D1)**:D1168–D1180 DOI 10.1093/nar/gkx1152.

Enciso-Rodriguez F, Douches D, Lopez-Cruz M, Coombs J, de Los Campos G. 2018. Genomic selection for late blight and common scab resistance in tetraploid potato (Solanum tuberosum). *G3 Genesgenetics* **8(7)**:2471–2481 DOI 10.1534/g3.118.200273.

FAO. 2022. FAOSTAT. Rome: food and agriculture organization of the united nations. *Available at http://www.fao.org/faostat/en/#data/*.

Fiorini N, Canese K, Starchenko G, Kireev E, Kim W, Miller V, Osipov M, Kholodov M, Ismagilov R, Mohan S, Ostell J, Lu Z. 2018. Best match: new relevance search for PubMed. *PLOS Biology* **16(8)**:e2005343 DOI 10.1371/journal.pbio.2005343.

Hernández LUH. 2014. Effect of systemic and contact fungicides on late blight disease and tuber yield of potato. *Journal of Agricultural Science & Technology* **10(1)**:209–217.

Hirsch CD, Hamilton JP, Childs KL, Cepela J, Crisovan E, Vaillancourt B, Hirsch CN, Habermann M, Neal B, Buell CR. 2014. Spud DB: a resource for mining sequences, genotypes, and phenotypes to accelerate potato breeding. *The Plant Genome* **7(1)**:365 DOI 10.3835/plantgenome2013.12.0042.

Holzschuher F, Peinl R. 2013. Performance of graph query languages: comparison of cypher, gremlin and native access in neo4j. In: *Proceedings of the Joint EDBT/ICDT 2013 Workshops*. 195–204.

Ivanisenko TV, Saik OV, Demenkov PS, Khlestkin VK, Khlestkina EK, Kolchanov NA, Ivanisenko VA. 2018. The solanum tuberosum knowledge base: the section on molecular-genetic regulation of metabolic pathways. *Vavilov Journal of Genetics and Breeding* **22(1)**:8–17 DOI 10.18699/VJ18.325.

Kondhare KR, Natarajan B, Banerjee AK. 2020. Molecular signals that govern tuber development in potato. *International Journal of Developmental Biology* **64(1-2-3)**:133–140 DOI 10.1387/ijdb.190132ab.

Larmande P, Do H, Wang Y. 2019. OryzaGP: rice gene and protein dataset for named-entity recognition. *Genomics & Informatics* **17(2)**:e17 DOI 10.5808/GI.2019.17.2.e17.

Majeed A, Muhammad Z, Ahmad H, Islam S, Ullah Z, Ullah R. 2017a. Late blight of potato (Phytophthora infestans) II: employing integrated approaches in late blight disease management. *PSM Biological Research* **2(3)**:117–123.

Majeed A, Muhammad Z, Ullah Z, Ullah R, Ahmad H. 2017b. Late blight of potato (Phytophthora infestans) I: fungicides application and associated challenges. *Turkish Journal of Agriculture-Food Science and Technology* **5(3)**:261–266 DOI 10.24925/turjaf.v5i3.261-266.1038.

Mausam M. 2016. Open information extraction systems and downstream applications. In: *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*. 4074–4077.

Meyer S, Nagel A, Gebhardt C. 2005. PoMaMo—a comprehensive database for potato genome data. *Nucleic Acids Research* **33(suppl_1)**:D666–D670 DOI 10.1093/nar/gki018.

Mihovilovich E, Sanetomo R, Hosaka K, Ordoñez B, Aponte M, Bonierbale M. 2015. Cytoplasmic diversity in potato breeding: case study from the International Potato Center. *Molecular Breeding* **35(6)**:1–10 DOI 10.1007/s11032-015-0326-1.

OpenAI. 2022. Chatgpt. *Available at https://chat.openai.com*.

Saik OV, Demenkov PS, Ivanisenko TV, Kolchanov NA, Ivanisenko VA. 2017. Development of methods for automatic extraction of knowledge from texts of scientific publications for the creation of a knowledge base Solanum TUBEROSUM. *Sel'skokhozyaistvennaya Biologiya [Agricultural Biology]* **52(1)**:1–74 DOI 10.15389/agrobiology.2017.1.63eng.

Schriml LM, Arze C, Nadendla S, Chang YWW, Mazaitis M, Felix V, Feng G, Kibbe WA. 2012. Disease Ontology: a backbone for disease semantic integration. *Nucleic Acids Research* **40(D1)**:D940–D946 DOI 10.1093/nar/gkr972.

Shrestha R, Matteis L, Skofic M, Portugal A, McLaren G, Hyman G, Arnaud E. 2012. Bridging the phenotypic and genetic data useful for integrated breeding through a data annotation using the Crop Ontology developed by the crop communities of practice. *Frontiers in Physiology* **3**:326 DOI 10.3389/fphys.2012.00326.

Singh G, Papoutsoglou EA, Keijts-Lalleman F, Vencheva B, Rice M, Visser RGF, Bachem CWB, Finkers R. 2021. Extracting knowledge networks from plant scientific literature: potato tuber flesh color as an exemplary trait. *BMC Plant Biology* **21(1)**:198 DOI 10.1186/s12870-021-02943-5.

Sun Q, Huang K, Yang X, Tong R, Zhang K, Poria S. 2024. Consistency guided knowledge retrieval and denoising in llms for zero-shot document-level relation triplet extraction. In: *Proceedings of the ACM on Web Conference*. 4407–4416.

Tiwari I, Shah KK, Tripathi S, Modi B, Subedi S, Shrestha J. 2021. Late blight of potato and its management through the application of different fungicides and organic amendments: a review. *Journal of Agriculture and Natural Resources* **4**:301–320 DOI 10.3126/janr.v4i1.33374.

Vleeshouwers VG, Finkers R, Budding D, Visser M, Jacobs MM, van Berloo R, Pel M, Champouret N, Bakker E, Krenek P, Rietman H, Jan Huigen D, Hoekstra R, Goverse A, Vosman B, Jacobsen E, Visser RG. 2011. Sol R gene: an online database to explore disease resistance genes in tuber-bearing Solanum species. *BMC Plant Biology* **11**:116 DOI 10.1186/1471-2229-11-116.

Wadhwa S, Amir S, Wallace BC. 2023. Revisiting relation extraction in the era of large language models. In: *Proceedings of the Conference. Association for Computational Linguistics. Meeting*, NIH Public Access, 15566.

Wei CH, Allot A, Leaman R, Lu Z. 2019. PubTator central: automated concept annotation for biomedical full text articles. *Nucleic Acids Research* **47(W1)**:W587–W593 DOI 10.1093/nar/gkz389.

Xing W, Qi J, Yuan X, Li L, Zhang X, Fu Y, Xiong S, Hu L, Peng J. 2018. A gene-phenotype relationship extraction pipeline from the biomedical literature using a representation learning approach. *Bioinformatics* **34(13)**:i386–i394 DOI 10.1093/bioinformatics/bty263.

Yuen J. 2021. Pathogens which threaten food security: *Phytophthora infestans*, the potato late blight pathogen. *Food Security* **13(2)**:247–253 DOI 10.1007/s12571-021-01141-3.