Overview of PestCLEF 2026: Knowledge Graph Extraction from Plant Health Literature

LifeCLEF at CLEF 2026

Robert Bossy[1,*], Marine Courtin[1], Louise Deléger[1], Xinzhi Yao[1,2] and Claire Nédellec[1]

[1]Université Paris-Saclay, INRAE, MaIAGE, Jouy-en-Josas, France
[2]College of Informatics, Hubei Key Laboratory of Agricultural Bioinformatics, Huazhong Agricultural University, Wuhan, China

Abstract

The first edition of PestCLEF, organized within LifeCLEF, challenges participants to extract knowledge graphs from Web documents in the domain of plant health. Understanding and monitoring crop disease transmission is essential for ensuring food security, economic stability, and environmental sustainability, yet relevant knowledge is dispersed across diverse sources that use varying terminology and perspectives. PestCLEF aims to advance large-scale, standardized extraction of crop disease knowledge from documents, enabling improved data integration, epidemiological surveillance, and cross-disciplinary understanding of plant disease mechanisms.

The PestCLEF data was built upon the EPOP corpus of 247 documents collected and annotated by the French plant epidemiomonitoring platform. The participants were required to extract the knowledge graph representing the domain knowledge and events, such as the cause of diseases, affected crops, or places and dates of occurrences of pest and vector species. PestCLEF pushes forward Information Extraction challenges with complex entities and long-range relations. The submissions were evaluated with the F-score computed by comparing the reference and predicted knowledge graph edges. The matching of edges in the evaluation process ensures that predicted edges are accounted for regardless of the node forms chosen by generative models.

Twelve teams participated in the PestCLEF challenge of which four outperformed the provided baseline. This paper describes 1) a detailed description of the challenge motivation, data, evaluation protocol, and baseline, 2) a description and analysis of the participating systems, and 3) a discussion on PestCLEF outcomes and future.

Keywords

Information Extraction, Knowledge Graph Extraction, Natural Language Processing, Document level Relation Extraction, Plant Health, Pest Control

1. Introduction

Protecting crops from pests and diseases requires access to reliable, up-to-date information with good spatial and temporal coverage so that experts can assess risks, anticipate outbreaks and make recommendations to limit pesticide usage. Currently, access to this information is available from large volumes of unstructured textual documents in the form of online media such as news alerts, monitoring reports and scientific publications which cannot be processed efficiently. Beyond these applications which address issues of food safety, environmental sustainability and economic perspective, there is also a more global interest in improving our understanding of ecological relations and species occurrence to enable policy-makers to address challenges in a context of climate change and biodiversity collapse.

Plant diseases, their pest agents, occurrences, and insect vectors form a complex network of knowledge scattered across a variety of sources. Due to different perspectives and target audiences (scientific or professional audience, with a more agronomic focus or a more ecological one), knowledge about plant diseases is expressed in diverse ways, using various vocabulary and registers. To enable large-scale analysis of these phenomena, it is necessary to integrate and aggregate data from various sources, to facilitate information extraction. The objective of the PestCLEF task is to develop accurate models for

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany
*Corresponding author.
Robert.Bossy@inrae.fr (R. Bossy); Marine.Courtin@inrae.fr (M. Courtin); Louise.Deleger@inrae.fr (L. Deléger); xinzhi_bioinfo@163.com (X. Yao); Claire.Nedellec@inrae.fr (C. Nédellec)
0000-0001-6652-9319 (R. Bossy); 0000-0003-4189-4322 (M. Courtin); 0000-0003-1399-4828 (L. Deléger); 0000-0001-6795-2653 (X. Yao); 0000-0002-0577-0595 (C. Nédellec)

large-scale knowledge extraction on crop diseases from documents. Such models would be invaluable for the cross-disciplinary and wide-range understanding of the mechanisms involved in crop diseases and the dynamics of their spread across ecosystems, which in turn will support the development of better epidemiological monitoring information systems.

From an information extraction perspective, PestCLEF corresponds to a knowledge graph extraction task, at the document-level. The task consists in providing a formal representation of the knowledge contained inside a document, in the form of a graph where nodes represent entities of interest, and where edges represent the relationships between these entities. It can be seen as a variant of the classical text-bound extraction task, where entities are anchored to the text through character or token offsets, and relations link entities without aggregation. This tradeoff, forgoing textual-boundedness and instead favoring an aggregated knowledge-representation at the document-level, was proposed to encourage participants to explore original methods regardless of their ability to provide entity offsets. Similar tasks have been proposed in the past, such as the knowledge base extraction subtask of Bacteria Biotope at BioNLP-ST 2016 [1] and BioNLP-OST 2019 [2] where the focus was on bacteria and their habitats. The knowledge graph extraction task is very similar to the document-level relation extraction (docRE) task like DocRED [3] because relations are not anchored to specific mentions. However docRE does not necessarily aggregate relations into a knowledge graph. In PestCLEF, the evaluation was carefully considered so that the matching of edges accounts for possible variations in node forms, as long as they are equivalent, which can be a penalizing factor when using generative models if not taken into consideration.

2. Task description

2.1. Dataset

The task is performed on the EPOP corpus [4], which is a collection of 247 web documents collected by the Plant Health Epidemiological Surveillance Platform (ESV Platform). These documents focus on 15 monitored pests and cover several text genres (news articles, professional journals and scientific literature). The original documents were automatically translated from 26 source languages into English by the ESV platform, and manually annotated by a team of 30 experts in epidemiology following state-of-the-art annotation methodology [5].

The annotation schema is designed to capture ecological interactions and occurrences in the context of complex epidemiological events. It involves text-bound annotations of named entities with character offsets of each mention, normalizations with references to shared resources, binary relations and n-ary relations. The EPOP annotation also features equivalence coreference links that gather entity mentions that refer to the same object or concept. Entity types are briefly described in Table 1.

For normalization, living organisms (typed as either *Pest*, *Vector* or *Plant*) are normalized using NCBI Taxonomy[6], while geographical locations are linked to their GeoNames[7] identifier. The original EPOP annotation was also enriched with disease identifiers. This plant disease lexicon was derived from the synonym lists of organism names in EPPO Global Database [8]. The construction of the lexicon involved automatic filtering to classify names as either pathogen names or disease names using surface clues and morpho-syntactic analysis, as well as manual curation to classify uncertain names. Each set of names sharing the same pathogens was then assigned a common identifier.

Ecological interactions and occurrences are annotated using binary relations between source and target entities. Figure 1 outlines the schema of PestCLEF knowledge graphs.

The annotations of each document from EPOP was transformed into a knowledge graph following a series of steps. First, all text-bound annotations of named entities involved in binary relations are considered as nodes of the graph. When entities either explicitly share the same reference identifier, or are linked by an equivalence coreference link, their nodes are collapsed. The surface forms of these text-bound annotations are collected in an array and constitute the acceptable forms for the node, which are deemed equivalent. The binary relations between the entities provide the edges of the graph, which are typed according to the relation type. If an edge is duplicated, that is if there are several occurrences

Entity type	Definition
Pest	Pathogen organism; always a microorganism; mentioned mostly by their scientific name (latin binomial).
Plant	Host crop like "citrus", "grape" or "wheat". Hosts are mentioned either by their scientific name, or by their vernacular name.
Disease	Disease caused by a pest. Disease names might be abbreviated.
Vector	Organism that transport pests and infect new crops. Vectors are mostly insects, but birds and bats might also vector diseases. Vectors are often mentioned by their scientific name.
Dissemination_pathway	Means of transportation of pests that contribute to their spread. Dissemination pathways are mostly plant parts (e.g. fruits), or human artifacts (e.g. crates).
Location	Geographical place where organisms or diseases are observed. Places might be of any scale (continent, country, county, municipality, etc.)
Date	Date, period or season when organisms or diseases are observed.

Table 1
Entity types of EPOP and their definitions and examples.

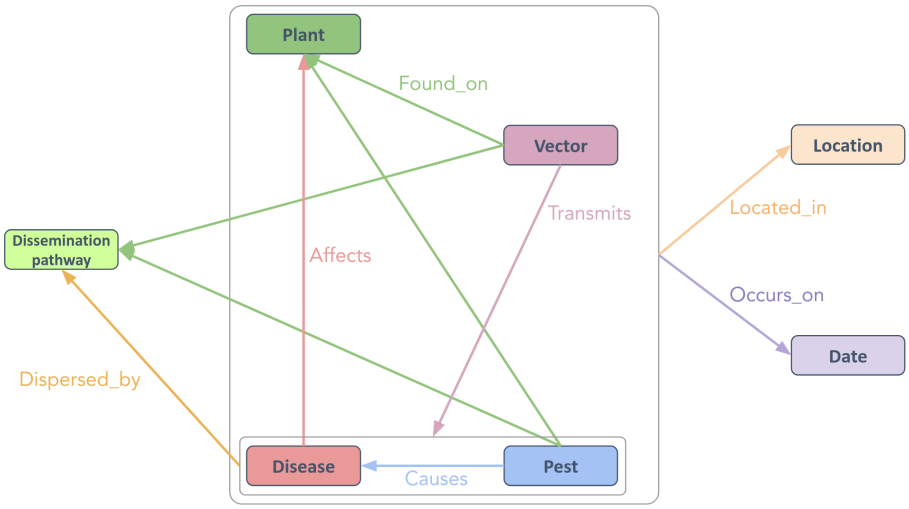

Figure 1: Schema and relation signatures for the PestCLEF knowledge graphs.

Edge type	Training	Development
Found_on	233	106
Transmits	25	5
Causes	44	27
Located_in	781	377
Occurs_on	209	95
Affects	50	30
Dispersed_by	32	16

Table 2
Frequency of edge types in the knowledge graphs of each subset of the official split.

of an edge with the same subject, object and predicate, only one is kept. Figure 2 showcases an example of this transformation process, for a short document excerpt. The distribution of edge types in the knowledge-graphs is unbalanced, as exemplified in Table 2.

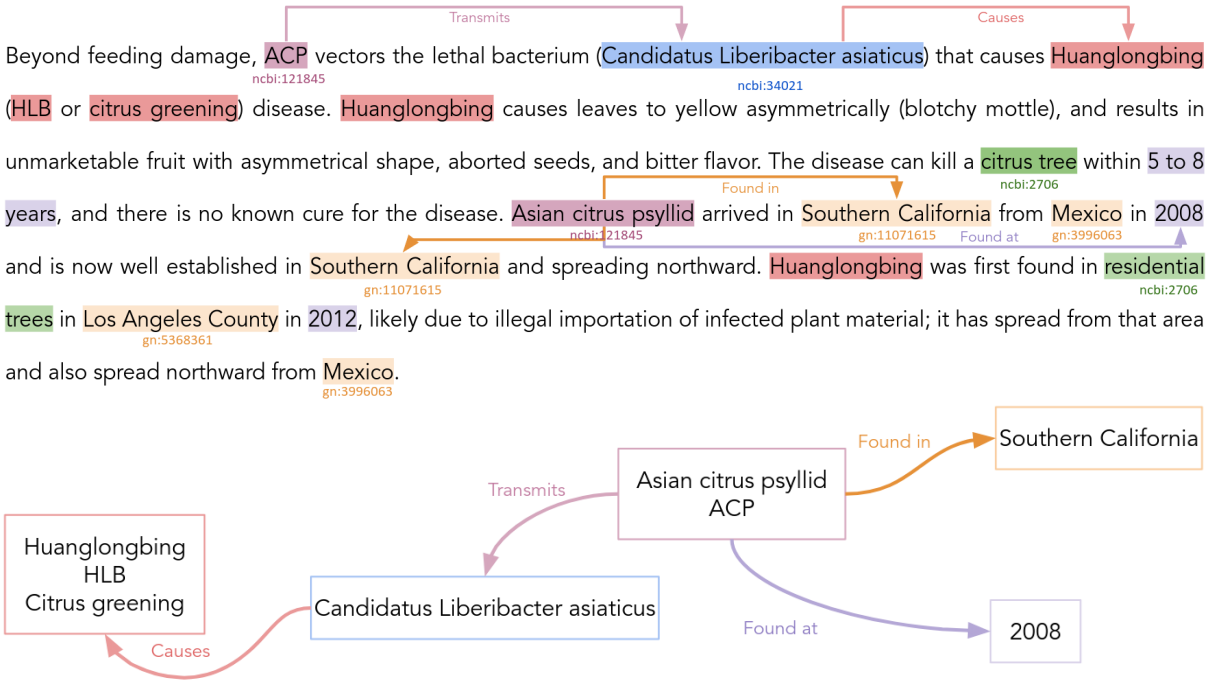

Figure 2: Sample paragraph with text-bound annotations and the corresponding knowledge graph. Annotations include named entities, normalizations and relations. Each vertex is associated with the set of forms that denote the corresponding entity.

2.2. Evaluation

PestCLEF aims at assessing the capabilities of current models and architectures at knowledge graph extraction at the document level. Thus, the evaluation of submissions relies on the comparison of an extracted knowledge graph with a ground truth knowledge graph. The comparison is edge-based because most of the knowledge contained in the documents are embodied in edges rather than nodes. Indeed nodes represent entities of the domain which are extensively defined in external resources. For instance, species whether they are host plants, pests or vectors, are registered in reference resources.

The evaluation measures for each document the similarity between the predicted and reference knowledge bases using the F_1 metric, as calculated below. The True Positives, False Negatives, and False Positives are counted from matching edges, edges missing from the prediction, and extra predicted edges respectively. The score for a set of documents is the average of the F-score computed on each document.

$$
\begin{aligned}
Recall &= \frac{TP}{TP + FP} \\
Precision &= \frac{TP}{TP + FN} \\
F_1 &= \frac{2 \times Recall \times Precision}{Recall \times Precision}
\end{aligned}
\tag{1}
$$

Participants were required to provide their submissions as a predicted knowledge base for each test document. The predicted knowledge base consists in a list of directed edges represented as triplets $\langle subject, predicate, object \rangle$. Where *subject* and *object* are the nodes connected by the edge, and *predicate* is the label of the edge. *subject* and *object* are strings representing the mentions of the connected entities in the text. *predicate* is a string chosen among the edge types in the PestCLEF schema. The evaluation of a knowledge graph extracted from a document is designed to take into account that the extraction operates at the document-level, thus the *subject* and *object* may refer to any equivalent

entity in the document. To this end, the evaluation process consists of three steps: edge matching, deduplication, then scoring.

2.2.1. Edge matching

The edge matching step associates, if possible, each edge from the predicted knowledge graph to an edge of the reference knowledge graph. A predicted edge E_p matches a reference edge E_r, if and only if:

- E_p *subject* is equal to one of the forms associated with E_r *subject*, and
- E_p *object* is equal to one of the forms associated with E_r *object*, and
- E_p *predicate* is equal to E_r *predicate*.

All string comparisons are case and whitespace insensitive, but strict with regards to inflection as no lemmatization or stemming is performed. This matching condition ensures that a predicted edge matches a reference edge regardless of the form the submission chose at both ends as long as they match one of the equivalent forms. This evaluation setup is similar to the one proposed by MAVEN-Arg event extraction [9] where event slots are counted correct if the prediction matches one of the mentions in a coreference chain. In PestCLEF, equivalent forms rely on the named entity normalization in addition of explicit equivalence coreference chains.

2.2.2. Deduplication

The deduplication step aims at removing edges that have been predicted several times. Indeed additional copies of the same edge in a predicted knowledge graph do not provide any useful information. Two predicted edges are considered equivalent if and only if both match the same reference edge. A single exemplar is retained in a set of equivalent edges, the others are discarded. The retained edge is chosen randomly as the choice does not affect the scoring. By this definition, predicted edges that do not match a reference edge cannot be discarded.

2.2.3. Scoring

The scoring step translates the matches produced by the previous step into True Positives, False Positives and False Negatives:

- each predicted edge that matches a reference edge (after deduplication) counts as a True Positive;
- each predicted edge that does not match any reference edge counts as a False Positive;
- each reference edge that is not matched by a predicted edge counts as a False Negative.

These counts allow us to compute the $Recall$, $Precision$ and F_1 for each document. F_1 cannot be computed in the case of documents for which the reference knowledge graph is empty. For these documents, a special scoring is applied:

- if the predicted knowledge graph is empty, then F_1 = 1.0;
- otherwise F_1 = 0.0.

2.3. Challenges

The dataset presents a number of challenges for information extraction methods. As with most specialized corpora, it contains domain-specific vocabulary, which may reduce the effectiveness of models trained on more general language data. Documents also come from a variety of online sources, including news articles, professional journals, and scientific literature, each characterized by different document lengths, registers and ways of expressing the information. For instance, scientific papers mainly refer to species using Latin taxonomic names whereas news articles more frequently employ vernacular

species names. Such heterogeneity makes patterns less predictable than in corpora focused on a single document type.

The multilingual nature of the dataset introduces an additional source of complexity. Documents were originally written in multiple languages and subsequently machine-translated into English, which can introduce noise such as mistranslations of specialized terminology.

Other challenges stem from the definition of the task and annotation scheme. Relations are annotated at the document level and frequently span multiple sentences. On average, relation arguments are separated by a distance of 20 words. Cross-sentence and long-range relations are a notoriously difficult issue in relation extraction, requiring extended context modeling. The annotation scheme also includes entity types that carry context-dependent semantic roles as well as intrinsic properties (e.g., *Dissemination_pathway*, *Pest* and *Vector*) and can be challenging to recognize compared to more traditional named entities such as locations and dates.

Finally, data scarcity is also an issue. Some entity and relation types have few positive examples, making them harder for a model to learn. This is especially the case for *Dissemination_pathway*and *Vector*entities, as well as for *Causes*, *Transmits*and *Dispersed_by*relations.

2.4. Baseline

A baseline prediction was provided in order to allow participants to assess the quality of their own predictions. A one-shot hard-prompt sent to a strong LLM was chosen because it would be the to-go method for a modern domain specialist. The prompt [10] sets the model as an information extraction system, defines the entity and relation schema with argument type constraints, and requires JSON output containing the extracted entities and relations.

The baseline used DeepSeek-V3 through the official DeepSeek API [11], accessed via the deepseek-chat model alias. The prompt was submitted through an OpenAI-compatible chat-completion interface.

The discrepancy between the published PestCLEF results and the scores reported in this shared task can be attributed to three factors: the baseline prompt does not cover the complete task schema; unparseable JSON outputs are handled differently, being excluded in [10] but assigned a score of zero in the official evaluation; and the two evaluations are conducted on different datasets, the development and test set respectively.

3. Challenge organization

3.1. Timeline

Figure 3 shows the timeline of the PestCLEF challenge. PestCLEF, along with all tasks in the LifeCLEF Lab [12], was announced in September 2025 at the CLEF 2025 conference. Until it was launched in January 2026, the task was advertised in mailing lists and social networks dedicated to Machine Learning and Natural Language Processing. From the 18th February, participants could access the full training dataset, and submit predictions on the test documents. The task description, data, submission, and evaluation were handled as a Kaggle challenge[1]. A Discord server was used as an instant messaging channel in order to ease the communication between the participants and organizers of all LifeCLEF tasks. The submissions were frozen the 7th May when the final results were disclosed. All participants were given the opportunity to submit until the 28th May a working notes paper describing their solution.

3.2. Dataset split

PestCLEF retained the official EPOP dataset split into training, development, and test sets. Participants were provided with the full training and development sets, including the document text, the original HTML tags, the gold knowledge base, and the original text-bound annotation. They were also provided

[1] https://www.kaggle.com/competitions/pest-clef-2026

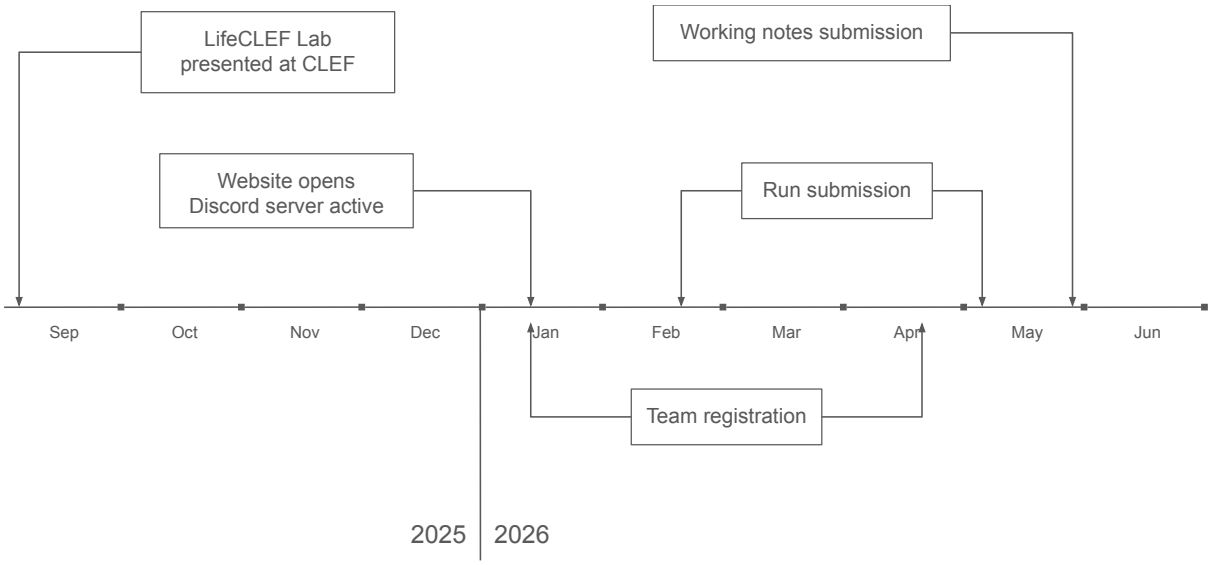

Figure 3: Timeline of the PestCLEF challenge.

Set	Documents	%
Training	110	44.53
Development	55	22.27
Private test	67	27.13
Public test	15	6.07
Total	247	100.00

Table 3
Official split of the PestCLEF dataset.

with the document text and HTML tags of the test set without the knowledge graph or text-bound annotation.

The test set was split between a private part and a public part as it is common practice on the Kaggle platform. During the challenge, participants could submit predictions on the test set which were evaluated on the public part. In this way, a public leaderboard was constantly available during the challenge while limiting the potential of overfitting participants' models. The two submissions of a participant that obtained the best scores on the public part were scored on the private part. The best of the two scores on the private was revealed at the end of the challenge.

The split was designed to be uniform: each set was as representative as possible of the whole corpus. A selection of variables were chosen to describe the corpus and subsets of the corpus. These include the total number of tokens, the original language, the number of entities of each type, and the number of relations of each type. The official split was chosen to minimize the sum of squared differences on these variables among a million randomly generated splits. Table (split) provides the size of each subset of the official split.

4. Results

4.1. Overview

PestCLEF received submissions from twelve teams who uploaded an average of 27 predictions each, although there is a large gap between seven teams who uploaded fewer than 10 predictions, and three

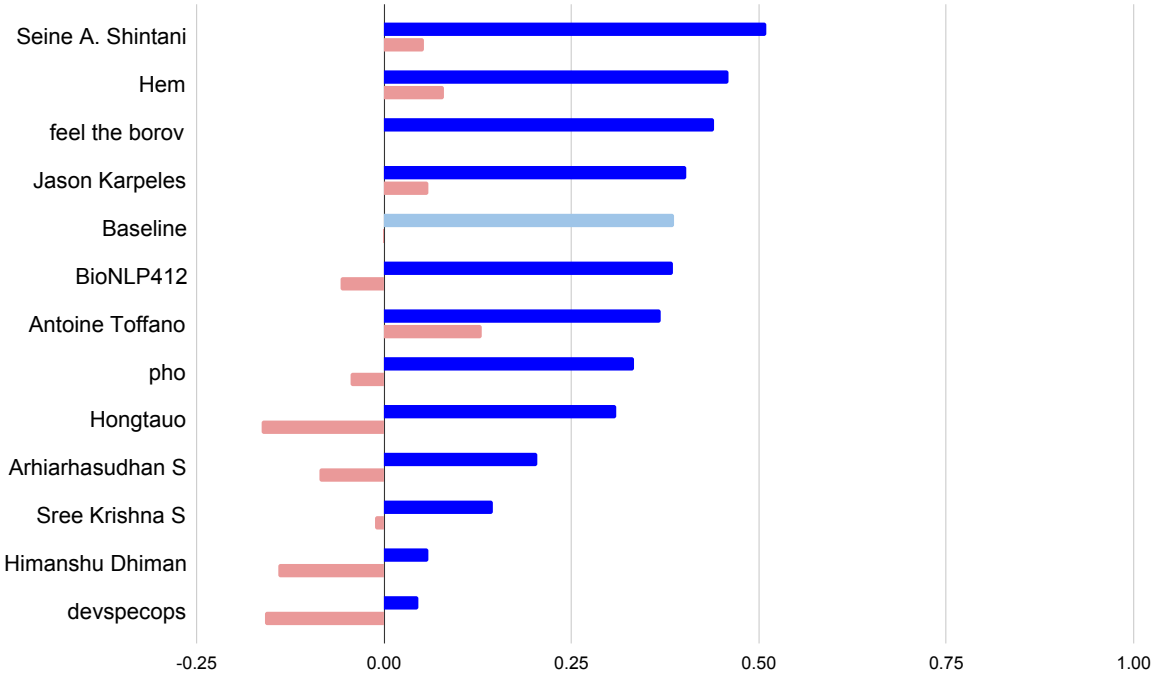

Figure 4: PestCLEF 2026 final results. Blue bars indicate the F1 score of their best submission on the private test set. Red bars indicate the score difference between the public and the private test set. The light blue bar indicates the performance of the baseline provided at the beginning of the challenge.

teams who uploaded more than 50. Figure 4 shows the final results along with the difference between the scores on the public and private test sets. Four teams outperformed the provided baseline and two more are within a bracket of 0.05 points below the baseline, which could be considered competitive. The performance drop from the public to the private test set is moderate for most participants, some even gaining a significant amount of points, although three submissions exhibit a large drop. The importance of the drop does not seem to be correlated with the amount of predictions submitted during the challenge. In overall, with the exception of a few teams, the proposed solutions do not seem to be exceedingly overfitting.

Figure 5 shows the scores of top-performing and publishing participants broken down by edge type. Unsurprisingly, abundant edges are better predicted: *Located_in*, *Found_on*, *Occurs_on*. Moreover these edges correspond to non-specific relations (location and time), one can hypothesize that argument entities are easier to detect and pretrained models already embed insights on these relations. However rarer edges, like *Causes*, *Affects*, and to a lesser extent *Transmits*, were well predicted. These edges also represent domain-specific knowledge that are particularly sought for.

The five notebook papers submitted by participants are summarized below. A clear pattern can be observed in the described systems. All participants report having implemented a classic information extraction pipeline, including named entity recognition, relation extraction, and post-processing. They used frugal techniques exploiting the text-bound annotations for training. Participants chose two main strategies for named entity recognition : either training a model from the BERT family, or mapping a lexicon acquired from external resources or the training set augmented with variation patterns. For relation extraction, all reporting participants trained a model for classifying relation candidates. Strategies differ in the generation of candidates, negative sampling, and the classifier algorithm. Post-processing steps were implemented in order to enforce schema constraints, remove duplicates, and correct erroneous predictions.

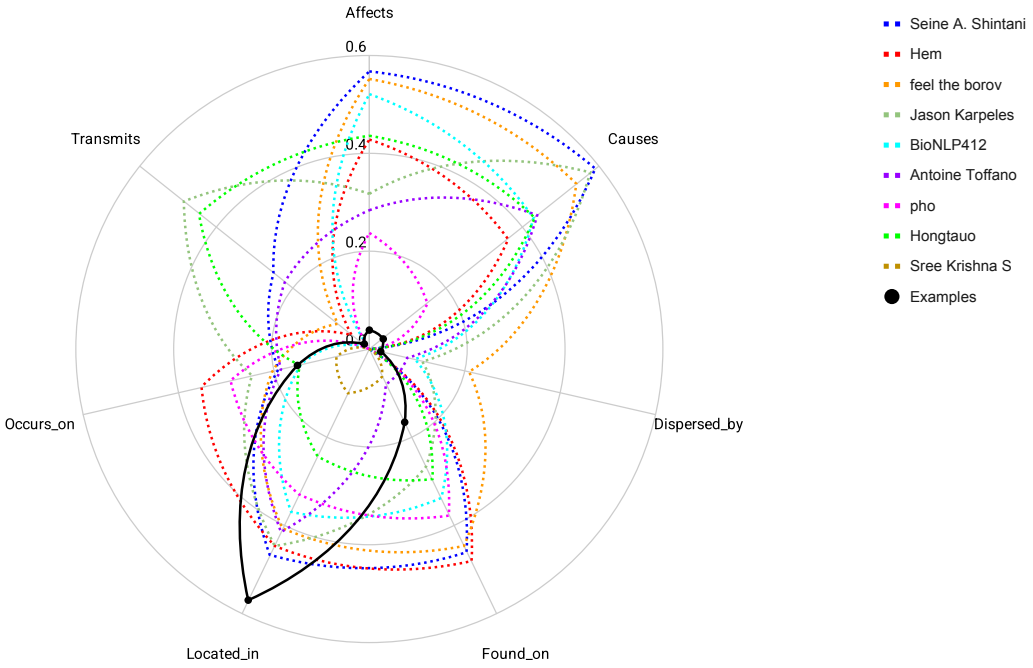

Figure 5: PestCLEF results by edge type. Only top-performing teams and teams that published are shown. Dotted lines represent the performance of each team on each edge type. The plain black line represents the relative abundance of edges of each type in the training and development sets.

4.2. Participating teams

4.2.1. Seine A. Shintani, Chubu University, Japan[13]

Team *Seine A. Shintani* ranked first with a F_1 of 0.51 on the private leaderboard, exceeding the baseline by 0.12. Their private leaderboard score exceeds their private leaderboard score by a little margin (0.03) showing that their solution is rather robust. Their solution is a hybrid machine-learning and hand-crafted rule-based system. The pipeline of *Seine A. Shintani* starts with a named-entity recognition by lexicon mapping augmented with regular expression expansions to handle common variations and abbreviations. The lexicon was built from entities provided in the training and development set. A second step identifies documents in the training set that are similar to the target document. The document similarity was computed from a TDF-IDF representation of the documents. The closest document is considered topically close to the target document. Edges from the closest training documents are transferred to the target document. This step allows the participants to short-circuit the prediction for trains of documents reporting very similar events. The final step classifies candidate edges generated from the edge type constraint signatures. The classifier is a decision tree trained by the LighGBM framework on the training set. *Seine A. Shintani* also went through several iterations of submissions on the public test set where each submission differed in the prediction for a single document. In this way they could refine their system by detecting probable sources of errors.

4.2.2. Hem, University of Bucharest, Romania[14]

Hem ranked third in the official evaluation achieving a F_1 of 0.46, outperforming the official baseline by 0.07 points and trailing the top-ranked system by 0.05 points. The proposed approach consists of a pipeline architecture of two BioMedBERT models fine-tuned on the EPOP dataset: the first performs NER while the second performs RE using the predicted named entities as input during training. Named entities are represented using the BIO tagging scheme. To mitigate the class imbalance in relation

extraction, the authors employ negative sampling. Predicted relations are subsequently merged using the coreference information available in the training data. Because the model input is limited to 512 tokens, long-distance relations cannot be captured. Precision is further improved by applying confidence thresholds at inference time.

The main strength of the method lies in the simplicity of its architecture combined with a small number of targeted adaptations specifically designed to address the characteristics of the EPOP dataset. In addition to its competitive performance, the approach remains computationally efficient, relying on lightweight fine-tuning of encoder models.

4.2.3. Feel the Borov, National Research Nuclear University MEPhI, Russia[15]

Team *Feel the borov* ranked 3rd in the challenge with an F_1 score of 0.44, outperforming the baseline by 0.05 points. The proposed approach combines dictionary-based entity recognition, classic supervised machine learning, and rule-based post-processing. Entity mentions are identified using a lexicon derived from the training set and enriched with entries from the NCBI Taxonomy and variants generated using heuristics. Candidate relations are generated within three sentences and classified using per-relation ensembles of GradientBoosting models trained on 26 handcrafted features, including lexical cues, TF–IDF metrics, argument distance, and corpus-derived statistics. Post-processing rules are then applied to remove likely false positives.

The main strengths of the method are its computational efficiency and interpretability. Unlike resource-intensive LLM-based approaches, it can be trained and executed on standard CPU hardware within minutes. Despite its simplicity, it achieved competitive results. Limitations include the entity recognition strategy that relies heavily on lexical resources derived from the training data, limiting its ability to generalize to unseen mentions, and the three-sentence candidate generation window that restricts the detection of long-range relations.

4.2.4. BioNLP412, University of Bucharest, Romania [16]

BioNLP412 obtained a F_1 of 0.39, ranked fifth, and outperformed the official baseline. The system used a two-stage ModernBERT-base pipeline fine-tuned on the text-bound annotations. In the first stage, ModernBERT-base was fine-tuned as a BIO token classifier to predict typed entity mentions, namely mention spans and their entity types. At inference time, the system used overlapping sequences of sentences, and the neural predictions were augmented with a mention lexicon built from the training annotations to improve entity coverage and precision. In the second stage, ModernBERT-base was fine-tuned for multi-label relation classification over schema-compatible subject-object candidates generated from the detected mentions. A schema-based post-processing step then removed invalid or duplicate triples so that the final output followed the task relation constraints. The best configuration used a five-seed agreement ensemble, retaining only candidate pairs predicted in at least four of the five runs. This reduced random-seed variance and filtered unstable mention-level false positives. The system's main strength is its controlled pipeline design, which combines task-specific fine-tuning, schema-aware candidate generation, post-processing, and conservative ensemble filtering.

4.2.5. Sree Krishna S, Sri Sivasubramaniya Nadar College of Engineering, India [17]

Team *Sree Krishna S* ranked 10th with a F_1 score of 0.1456, which is 0.2419 below the baseline method. The approach is a pipeline system with three components : a token-level classification model for named entity recognition which fine-tunes BioBERT, a candidate entity pair classification model which leverages entity markers for relation extraction and a module for aggregating the predicted relations into a knowledge graph based on a confidence threshold. For relation extraction, the dataset is augmented with negative examples constructed using schema-valid entity pairs appearing inside the same document in close proximity, for which there is no relation. To reduce the effect of class imbalance, weighted cross entropy is used as a loss for both named entity recognition and relation extraction. The proposed

method performs very similarly on the development set and test set, indicating that it generalizes well and does not overfit the data.

5. Conclusions and Discussion

The PestCLEF challenge was built to support both end-to-end models and classic step-by-step pipelines that make use of intermediate, text-bound annotations. In practice, most teams chose the traditional pipeline route, using the provided annotations for supervision; these approaches generally outperformed –or at least matched– the zero-shot end-to-end baseline. Because the solutions were also computationally lightweight, extracting knowledge graphs from large corpora becomes feasible. This outcome demonstrates that well-annotated corpora remain valuable for training robust and scalable information-extraction systems, even as zero- and few-shot methods continue to improve and become more efficient. Future versions of PestCLEF could broaden the range of methods by defining separate tasks for knowledge-graph extraction with and without intermediate annotations.

The current evaluation was heavily dominated by generic edges describing location and time, which eclipsed more specialized, domain-specific information. A more balanced assessment would give a clearer picture of the practical usefulness of the proposed systems. Moreover the split was designed for uniformity between the training, development and both test sets. While this setup evaluates fairly systems in the first edition of a competition, it hinders the appraisal of generalization of proposed solutions in a production context where the stream of documents might exhibit a focus shift and the emergence of novel knowledge. Future editions might shift from a balanced split to a chronological, or even an adversarial split.

PestCLEF marks the first information-extraction task focused on plant health, opening the way for tools that can help plant scientists and protection agencies tackle major environmental challenges, and it highlights the diversity and promise of current IE technologies.

Acknowledgments

This work was supported by government funding managed by the French National Research Agency (ANR) under the France 2030 plan (EcoControl project, reference ANR-24-PEAE-0004, within the PEPR Agroecology and Digital).

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT 5.5 for translation and rephrasing. After using this service, the authors reviewed and edited the content as needed. The authors take full responsibility for the publication's content.

References

[1] L. Deléger, R. Bossy, E. Chaix, M. Ba, A. Ferré, P. Bessières, C. Nédellec, Overview of the bacteria biotope task at BioNLP shared task 2016, in: C. Nèdellec, R. Bossy, J.-D. Kim (Eds.), Proceedings of the 4th BioNLP Shared Task Workshop, Association for Computational Linguistics, Berlin, Germany, 2016, pp. 12–22. URL: https://aclanthology.org/W16-3002/. doi:10.18653/v1/W16-3002.

[2] R. Bossy, L. Deléger, E. Chaix, M. Ba, C. Nédellec, Bacteria biotope at BioNLP open shared tasks 2019, in: J.-D. Kim, C. Nédellec, R. Bossy, L. Deléger (Eds.), Proceedings of the 5th Workshop on BioNLP Open Shared Tasks, Association for Computational Linguistics, Hong Kong, China, 2019, pp. 121–131. URL: https://aclanthology.org/D19-5719/. doi:10.18653/v1/D19-5719.

[3] Y. Yao, D. Ye, P. Li, X. Han, Y. Lin, Z. Liu, Z. Liu, L. Huang, J. Zhou, M. Sun, DocRED: A large-scale document-level relation extraction dataset, in: A. Korhonen, D. Traum, L. Màrquez

(Eds.), Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 764–777. URL: https://aclanthology.org/P19-1074/. doi:10.18653/v1/P19-1074.

[4] C. Nédellec, M. Courtin, X. Yao, M. Grosdidier, I. Pieretti, S. Duperier, R. Bossy, EPOP: A Benchmark Corpus for Assessing NLP Models on Structured Information Extraction in Plant Health, in: S. Piperidis, N. Bel, H. van den Heuvel, N. Ide, S. Krek, A. Toral (Eds.), Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), European Language Resources Association (ELRA), Palma, Mallorca, Spain, 2026, pp. 1331–1340. doi:10.63317/4in2fpefq4pz.

[5] C. Grouin, T. Lavergne, A. Névéol, Optimizing annotation efforts to build reliable annotated corpora for training statistical models, in: L. Levin, M. Stede (Eds.), Proceedings of LAW VIII - The 8th Linguistic Annotation Workshop, Association for Computational Linguistics and Dublin City University, Dublin, Ireland, 2014, pp. 54–58. URL: https://aclanthology.org/W14-4907/. doi:10.3115/v1/W14-4907.

[6] C. L. Schoch, S. Ciufo, M. Domrachev, C. L. Hotton, S. Kannan, R. Khovanskaya, D. Leipe, R. Mcveigh, K. O'Neill, B. Robbertse, S. Sharma, V. Soussov, J. P. Sullivan, L. Sun, S. Turner, I. Karsch-Mizrachi, NCBI Taxonomy: a comprehensive update on curation, resources and tools, Database 2020 (2020) baaa062. doi:10.1093/database/baaa062.

[7] Wick, Mark, GeoNames, 2026. URL: https://www.geonames.org/.

[8] EPPO, EPPO Global Database, ???? URL: https://gd.eppo.int/.

[9] X. Wang, H. Peng, Y. Guan, K. Zeng, J. Chen, L. Hou, X. Han, Y. Lin, Z. Liu, R. Xie, J. Zhou, J. Li, MAVEN-ARG: Completing the puzzle of all-in-one event understanding dataset with event argument annotation, in: L.-W. Ku, A. Martins, V. Srikumar (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand, 2024, pp. 4072–4091.

[10] X. Yao, C. Nédellec, J. Xia, R. Bossy, Consistency–accuracy correlation in hard-prompted LLMs for entity and relation extraction: empirical findings from plant-health data, Genomics & Informatics 24 (2026) 3.

[11] DeepSeek-AI, Deepseek-v3 technical report, 2024. URL: https://arxiv.org/abs/2412.19437. arXiv:2412.19437.

[12] L. Picek, L. Adam, S. Kahl, R. Bossy, L. Chrobak, H. Goëau, K. Papafitsoros, H. Klinck, W.-P. Vellinga, R. Planqué, T. Denton, K. Barnard, C. Nédellec, L. Deléger, M. Courtin, G. Martellucci, I. Moummad, F. Vinatier, P. Bonnet, A. Joly, Overview of LifeCLEF 2026: Ai challenges for biodiversity understanding and ecosystem management, in: International Conference of the Cross-Language Evaluation Forum for European Languages (CLEF), Springer, 2026.

[13] S. A. Shintani, Hybrid Retrieval-Guided Knowledge Graph Extraction with Precision Repair for PestCLEF 2026, in: Working Notes of CLEF 2026 - Conference and Labs of the Evaluation Forum, 2026.

[14] A.-A. Hemeci, Knowledge Graph Extraction from Agricultural Pest Documents: A BiomedBERT Pipeline with NER and Relation, in: Working Notes of CLEF 2026 - Conference and Labs of the Evaluation Forum, 2026.

[15] L. Kokovkin, Lexicon Engineering and Gradient Boosting for Plant Pest Relation Extraction, in: Working Notes of CLEF 2026 - Conference and Labs of the Evaluation Forum, 2026.

[16] M. R. Radulescu, T. G. Arabagiu, BioNLP412: Knowledge Graph Extraction on Plant Pests from Web Documents, in: Working Notes of CLEF 2026 - Conference and Labs of the Evaluation Forum, 2026.

[17] S. Sree Krishna, K. J. Varghese, M. Sujith, B. Prabavathy, Sree Krishna S at PestCLEF 2026: Knowledge Graph Extraction from Plant Pest Documents Using BioBERT-based NER and Marker-based Relation Classification, in: Working Notes of CLEF 2026 - Conference and Labs of the Evaluation Forum, 2026.