EPOP: A benchmark corpus for Assessing NLP Models on Structured Information Extraction in Plant Health

Claire Nedellec[1], Marine Courtin[1], Xinzhi Yao[1,2], Marie Grosdidier[3], Isabelle Pieretti[4], Sandy Duperier[3], Robert Bossy[1]

[1]Paris-Saclay University, INRAE, MaIAGE, Jouy-en-Josas, France
[2]Huazhong Agricultural University, College of Informatics, Hubei Key Laboratory of Agricultural Bioinformatics, Wuhan, China
[3]ESV Platform, INRAE, Biostatistics and Spatial Processes unit (BioSP), Avignon, France
[4]ESV Platform, CIRAD, UMR PHIM, F-34398 Montpellier, France
{claire.nedellec, marine.courtin, xinzhi.yao, marie.grosdidier, sandy.duperier, robert.bossy}@inrae.fr , xinzhi_bioinfo@163.com, isabelle.pieretti@cirad.fr

Abstract

We introduce the EPOP (*Epidemiomonitoring of Plants*) corpus, a new annotated resource for structured information extraction in the domain of plant health epidemiology. The corpus consists of translated news reports that reflect real-world phytosanitary monitoring scenarios. It includes annotations for named entities (e.g. Plant, Pest, Vector, Disease, Dissemination Pathway), identity coreferences, and both binary and complex n-ary relations that represent key events such as Transmits or Causes, along with their modalities. A distinctive feature of EPOP is its normalization layer where mentions of species and geographical locations are linked to canonical identifiers in the NCBI Taxonomy and GeoNames, enabling semantic disambiguation and integration with external knowledge bases. As the first publicly available corpus of its kind, EPOP presents a realistic and challenging benchmark, with high linguistic variability, entity role ambiguity, and long-distance relations. We report baseline results on core tasks (named entity recognition, normalization (entity-linking), and relation extraction) using both fine-tuned BERT-based models and hard-prompted large language models. These experiments demonstrate the utility of EPOP while also identifying areas for improvement, particularly in the extraction of complex relations. The corpus is released under an open license, to support research in environmental NLP, crop protection, and knowledge graph enrichment.

Keywords: plant health epidemiology, trophic relation, annotated corpus, structured information extraction, n-ary relation extraction, entity normalization (linking), benchmarking.

1. Introduction

Protecting crop health from pests and disease threats requires real-time access to reliable information extracted from unstructured text in online media such as news alerts, monitoring reports and scientific communications. However, the field of plant epidemiology remains underexplored in NLP, largely due to the absence of specialized, annotated resources that reflect its complexity.

Crop health surveillance presents distinct challenges in NLP, including overlapping events and semantic ambiguity inherent to real-world complexity. To enable actionable decision-making, extracted information must be both semantically grounded and verifiable. This requires normalization of entities to canonical identifiers (e.g., NCBI Taxonomy for species, GeoNames for locations) for integration with knowledge graphs and predictive models. The extracted information must also be easily verifiable by the monitoring platform end-users, which calls for text-bound annotations that allow direct visual inspection of highlighted entities within source documents.

To address these needs, we present the Epidemiomonitoring of Plants (EPOP) corpus, the first publicly available resource designed for structured information extraction in plant health epidemiology. The source documents were collected from the international health monitoring system of the French plant health epidemiological surveillance platform (ESV Platform), which compiles multilingual news, expert bulletins, and scientific reports to produce surveillance alerts and summary reports. EPOP is built with a rich annotation schema covering discontinuous entities, identity coreference, n-ary relations (events), and normalization of species and geographical locations. It supports the modeling of complex epidemiological scenarios such as new occurrences of pests, disease emergence and biological transmission pathways. EPOP includes span-level entity annotations to support the development of monitoring systems that can highlight mentions in context.

A distinctive feature of EPOP is its annotation of relation modality, which captures whether a relation is negated, hypothetical, or uncertain (e.g., “X was suspected”, “Y has not been confirmed”). This is an essential feature in

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 1331–1340
11–16 May 2026.

phytosanitary reports for assessing evidential reliability. To further support event-level extraction, EPOP provides identity coreference annotations that link mentions of the same entity in the same role across the document.

We report benchmark results on three core tasks, named entity recognition, normalization, and relation extraction using both fine-tuned small models and prompted large language models. By releasing EPOP publicly, we aim to advance NLP methods for plant health monitoring and contribute to broader efforts in environmental text mining and the acquisition of interoperable knowledge.

2. Related Work

In the plant health domain, the corpora for the NER tasks are small and focus on specific crops. They usually include crops, varieties, diseases, pathogen agents and plant parts. Examples include *KIWID* on kiwi plant (Zhang et al., 2022), AwdpCNER on wheat and corn, and ApdCNER on apple (Zhang et al. 2021). Some corpora also target limited geographical areas such as FloraNER for New Caledonia plant flora (Nainia et al., 2024) and COPIOUS on Philippines biodiversity (Nguyen et al. 2019).

To effectively report outbreak events and enable downstream integration, however, more structured representations are required. Substantial advances in structured information extraction have been driven by annotated corpora from the biomedical domain, particularly through initiatives like the BioNLP Shared Tasks series (Kim et al., 2019) and the BioCreative challenges (Miranda-Escalada et al., 2023). Beyond named-entity recognition, these efforts have established the importance of extracting relationships and events between entities and linking them to standardized terminologies and ontologies. The CRAFT corpus (Bada et al., 2012) exemplifies it with normalization of biomedical concepts across multiple ontologies, with rich cross-document annotations. Most of the public corpora with entity normalizations are related to biomedicine, although some corpora also include plant and pest species. The Bacteria Biotope (BB) corpus pioneered ontology-based entity normalization and used the OntoBiotope ontology for microbe habitats and the NCBI Taxonomy for species (Bossy et al., 2019). This methodology is also exemplified in other recent resources, such as the S1000 corpus for biomedical literature (Luoma et al., 2023) and the Taec corpus for crop traits (Nédellec et al., 2024), both of which perform species normalization In Biodiversity research, COPIOUS also includes useful geolocation normalization by coordinates which is rare in the plant domain though frequent in animal and human health.

Only a few plant-related corpora extend beyond entities to include relationships and events. Examples include Bacteria Biotope (Bossy et al., 2019) and COPIOUS (Nguyen et al. 2019) that link species to their habitats, corpora from molecular biology such as SeeDev (Chaix et al., 2016) on seed development, PICKLE (Lotreck et al, 2024), and (Singh et al, 2021) on genetic-phenotype associations. Yuan et al. (2024) introduced a corpus of Chinese web pages on cotton pests and diseases where the relationships include Control methods and Transmission path which are highly relevant for plant health monitoring. However, disease and pest mentions are merged under a single entity type and the annotations conflate taxonomic and lexical relations, which limits reuse.

Most of these corpora are derived from scientific literature and their entity and relation schemas do not align with the need for real-time event alert detection. Research on modeling spatiotemporal outbreak events in text has largely focused on document classification from social media or news sources for human and animal health (Kim et al., 2020), with only limited applications to crop monitoring (Shankar et al., 2020).

While modality is common in phytosanitary texts, it remains largely unaddressed in plant-related corpora. To our knowledge, EPOP is the first corpus in plant health to annotate relation modality (e.g., hypothetical, negated), capturing the evidential status of event mentions.

To fill this gap, we introduce the EPOP corpus, a collection of news reports that reflect real-time phytosanitary monitoring with the aim to model the underlying biological and trophic relations that drive epidemic emergence. EPOP model defines a comprehensive set of entities specific to plant health (e.g., Disease, Vector, Host Plant) and introduces structured event to represent complex transmission and environmental factors. It supports event structures with multiple arguments such as *vector*, *pathogen*, *host*, *location*, and *date* thereby mirroring real-world epidemiological processes.

Finally, while standard in general-domain NLP and present in biomedical corpora such as CRAFT (Bada et al., 2012), EPOP introduces identity coreference annotations specifically designed for epidemiological knowledge graph construction, adapting the approach used in Bacteria Biotope corpus to the plant health domain.

3. Corpus Construction

3.1 Data Collection

The source documents for the EPOP corpus were collected from the document repository of the international health monitoring of the ESV Platform that aggregates news articles, professional journals, and scientific literature to produce weekly and monthly surveillance reports. The corpus focuses on 15 high-priority monitored species, including bacteria, viruses, fungi and insects (e.g. *Xylella fastidiosa*, *Popillia japonica*, Tomato brown rugose fruit virus).

From an initial backlog of 100,000 documents, we selected a pool of the 1000 documents last processed by the ESV Platform. The recency of documents ensures that annotated documents reflect the most up-to-date practices of the ESV Platform. Documents were gradually sampled from this pool for the annotation according to experts availability and species representativeness. All documents were translated by the ESV Platform from 26 different languages to English using Google Translate and cleaned using the Trafilatura Python library (Barberesi, 2021) to extract plain text content.

3.2 Annotation Schema

The EPOP formal annotation schema is designed to capture complex epidemiological events and support NLP model training and evaluation. It comprises four annotation layers that provide a formal structured representation of the unstructured text.

We define a set of seven types of entity central to plant epidemiology: Pest, Plant, Disease, Pest Vector, Dissemination pathway (trade, wind, human activity), Location and Date.

Relations are defined with strict type constraints and include both binary and n-ary events. Binary relations are

Found_on: Pest|*Vector* → *Plant*

Transmits: *Vector* → *Pest*

Causes: *Pest* → *Disease*

Located_in: *Disease*|*Pest*|*Plant*|*Vector* → *Location*

Detected_on: *Disease*|*Pest*|*Plant*|*Vector* → *Date*

Affects: *Disease* → *Plant*

Dispersed_by: Disease|*Pest* → *Dissemination_pathway*

N-ary events represent either a complex knowledge bit, or an observation. They link multiple arguments with semantic roles: *Causes*, *Affects*, *Transmitted_by*, *Dispersed_by*, *Located_in*, *Detected_on*. Event arity ranges from 2 to 6.

Furthermore, identity coreferences link entity mentions of the same type that reference the same subject. A coreference asserts that entities are equivalent arguments in relations or events.

The pest, vector and plant entities represent living organisms and are linked to their corresponding NCBI Taxonomy identifiers, while geographical location entities are mapped to GeoNames ID. Both NCBI taxonomy and GeoNames serve as authoritative references that enable semantic integration with external databases. The hierarchical organization of these resources enables multi-level indexing, allowing annotations to capture varying degrees of specificity and precision. This flexibility is particularly useful for integrating heterogeneous sources and supporting downstream tasks such as reasoning, aggregation, or approximate matching in knowledge graphs.

Two relation modalities capture the evidential status of event mentions: negation and hypothesis.

We developed the EPOP annotation guidelines document to serve as the reference manual for all annotators to provide detailed instructions, definitions and examples for consistent and reproducible annotation. It is intended to support annotator training, quality control and LLM prompting. It was refined iteratively through annotation phases and annotator feedback.

3.3 Annotation Process

The annotation of the EPOP corpus was conducted through a structured, multi-phase process. Each document was first automatically annotated by a straightforward dictionary look-up and then annotated by trained annotators, experts in plant health in a double-blind way, followed by an adjudication phase to resolve all disagreements. The annotation guidelines were continuously refined based on recurring disputes and edge cases discovered during this adjudication stage. We then apply a set of patterns for the automatic detection and report of common annotation errors that require little expertise to correct. These warnings were checked by a single independent annotator.

Manual annotations were performed by 30 experts in epidemiology using the AlvisAE editor (Papazian et al., 2012). AlvisAE provides a collaborative user-friendly interface for complex annotation tasks including discontinuous, overlapping entities, events and ontology-based tagging and coreference. It supports annotator management and adjudication of conflicting annotations.

Inter-annotator agreement was measured by computing F_1 scores per document between the two annotators, then averaging across all shared documents. For named entities, the average F_1 is 0.75, and for normalizations the average F_1 is 0.70, indicating moderate consistency. Given the complexity of the domain and annotation schema, these values reflect the inherent ambiguity in defining entity boundaries, types, and event arguments.

3.4 Corpus characteristics

The resulting dataset comprises a total of 247 documents. Table 1 gives general statistics.

Token	Entity	Binary relation	N-ary relation	Coref.
115,000	7,537	4,717	2,929	373

Table 1: EPOP dataset statistics

We split the EPOP corpus into 3 parts, training (45%), development (22%) and test (33%). The test part remains hidden. Table 2 and 3 gives the distribution of the entity and relation annotations per type.

Entity type	Training	Dev
Date	419	217
Disease	234	148
Dissemination_pathway	138	48
Location	1042	485
Pest	908	338
Plant	663	347
Vector	78	32
Total	**2925**	**1350**

Table 2: Number of entities per type in EPOP training and development (Dev) datasets

Relation type	Training	Dev
Causes	66	35
Detected_on	287	134
Dispersed_by	36	18
Affects	141	74
Found_On	441	181
Located_In	1210	567
Transmits	36	13
Total	1894	870

Table 3: Number of binary relations in EPOP training and development (Dev) datasets

The average distance between relation arguments is 20 words, indicating that most relations occur across sentence boundaries and require modeling long-range context.

We observed that mentions of the same entity may play an equivalent role as arguments of relations which were not easily distinguishable and the distinction was not relevant with respect to plant epidemiology monitoring. We defined identity coreference sets to represent this equivalence. Figure 1 shows an example where two relations are relevant although the one figured by the solid line is better. The *Fusarium oxysporum f. sp. Cubense fungi* and *Foc* entities play the same semantic role as the agent of the *furiosis* disease and are therefore grouped within the same identity coreference set.

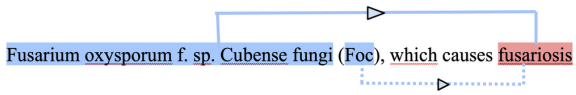

Figure 1. Example of identity coreference

The distribution of taxonomic and geographic identifiers in the normalization annotations highlights the broad diversity of pests and locations, with 138 distinct species of pests, many more than the initially targeted ones, and 459 unique geographic locations. The high frequency of the plant pathogen *Xylella fastidiosa* in Italy and Southern France reflects its prominence during the period of corpus compilation. Figure 2 shows a screenshot of an example of manual annotation of the pinewood nematode dissemination with interlinked events.

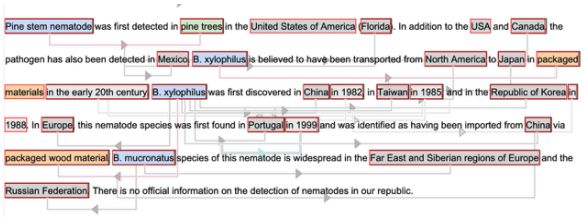

Figure 2. Example of event annotations on pinewood nematode using AlvisAE

Figure 3 gives an example of an HLB (Huánglóngbìng) contamination of citrus crop in California annotated with events and normalized identifiers.

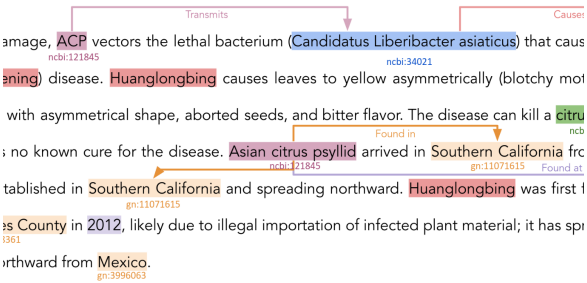

Figure 3. Example of HLB annotations, events and entity linked to identifiers

4. Benchmark Experiments

4.1 Objectives

To assess the utility of the EPOP corpus for structured information extraction in plant health, we define benchmark tasks aligned with its annotation layers: Named Entity Recognition (NER) and Relation Extraction. We provide baseline results using both fine-tuned small models and hard-prompted large language models, enabling comparison across modeling methods. We adapted evaluation metrics to reflect both general NLP performance and domain-specific needs. Notably, identity coreferences are used for event-level evaluation, and semantic distance in taxonomic is applied to assess normalization accuracy in a similar ways as (Bossy et al., 2012).

4.2 Small Models: Fine-tuning BERT and ReBERT

For the NER task, we implemented an architecture that consists of BioBERT (Lee et al., 2020), a transformer pretrained on a biomedical corpus, a softmax layer to classify tokens and an entity span reconstruction step. We trained a single model that predicts all entity types.

We achieved binary relation extraction ReBERT (Tang, 2023). ReBERT classifies relation candidates as either one of the relation types, or the absence of relation using the representation by Zhong and Chen (2021). The classification of a relation candidate is implemented as a BERT sequence classification through the embeddings of the [CLS] token. Each relation candidate is represented as the text where the boundaries of the candidate arguments are marked with special tokens (“@@” and “$$” respectively for the head and tail).

4.3 Large Language Models: Hard Prompting

We conducted our experiments with large language models (LLMs) using hard prompting to assess their one-shot capabilities without any additional fine-tuning.This approach aligns with the plant health domain where training data is scarce. Hard prompts allow us to directly query pretrained models using a template that mirrors the rich annotation schema of EPOP.

However, LLM effectiveness in event extraction may be limited by the strong requirements on semantic constraints (e.g., correct roles of arguments) and syntactic structure (e.g., well-formed JSON). We experimented with four widely used LLMs, GPT-4o-mini, Kimi, DeepSeek-V3, Qwen3. All models were accessed through OpenAI's Python API.

The models were prompted to perform NER and relation extraction in a one-shot setup, as splitting the task into two steps resulted in a significant drop in relation extraction accuracy. Relations are represented in JSON as triplets ⟨source, type, target⟩, where *source* and *target* refer to the surface forms of the relation arguments, and *type* denotes the relation label.

We applied a post-processing step to correct format errors in malformed JSON output (*e.g.* extra commas, unmatched delimiters). Outputs that remain invalid after correction are excluded.

We evaluate at the document level (DocRE) to bypass the limitations of LLMs in accurately generating character-level offsets for named entities. This choice reflects a trade-off between model accessibility and annotation precision. However, we acknowledge that this simplification does not address the practical needs of crop health monitoring, where users require precise graphical overlays of entities and relations within the text to efficiently review extracted information. For the DocRE task, reference annotations were constructed by merging strictly identical events across the documents, with identity coreference sets taken into account.

4.4 Results and Comparison

BioBERT models were fine-tuned for NER on the training and development sets of EPOP and evaluated on the test set with strict boundary matching. For hyperparameters, we set the batch size to 16, use a learning rate of 2e−5, and train the model for a maximum of 15 epochs with early stopping triggered if the cross-entropy loss does not improve for 5 consecutive epochs. We report Recall, Precision and F_1 performances in Table 4. Mean and standard deviations were computed with five seeds.

Entity	F_1	Recall	Precision
Any	0.81±0.01 (0.07)	0.84±0.01 (0.08)	0.78±0.01 (0.07)
Date	0.79±0.01 (0.10)	0.82±0.01 (0.10)	0.75±0.01 (0.09)

Disease	0.87±0.02 (0.05)	0.90±0.02 (0.05)	0.84±0.02 (0.05)
Dissemination_pathway	0.49±0.04 (0.17)	0.52±0.04 (0.18)	0.49±0.04 (0.16)
Location	0.84±0.01 (0.07)	0.85±0.01 (0.08)	0.83±0.02 (0.07)
Pest	0.85±0.01 (0.04)	0.91±0.01 (0.05)	0.79±0.01 (0.04)
Plant	0.79±0.01 (0.09)	0.83±0.01 (0.09)	0.76±0.02 (0.09)
Vector	0.36±0.05 (0.13)	0.32±0.04 (0.11)	0.45±0.07 (0.15)

Table 4. Performance of BioBERT in the NER task in the EPOP corpus. Difference between strict and relaxed measure into parenthesis.

These results show that lightweight domain-adapted models like BioBERT can achieve strong performance (F_1 > 0.80) when provided with sufficient and well-defined training data. Notably, the lower scores for Dissemination_pathway and Vector may be attributed to data scarcity since these two types have fewer than 200 annotated mentions in the training set, compared to over 300 for most other types and up to 1,500 for Location. The performance on Dissemination_pathway may also be affected by its context-dependent nature, as this type represents a semantic role rather than an independent category, unlike Plant for instance. Precision and recall are generally balanced across types with a slight skew toward recall.

For the relation extraction task, ReBERT was trained with a batch size of 32 and a learning rate of 2e-5 for up to 30 epochs with an early stopping triggered if three consecutive epochs do not lower the loss.

Candidate pairs are generated from gold-standard entities that satisfy schema constraints and are located either within the same sentence, across two consecutive sentences, or between a section title and any sentence within the same section. We chose this candidate generation heuristic to capture cross-sentence relations while keeping the number of candidates limited. We report the evaluation using Recall, Precision and F_1 in Table 5. Results are reported as mean and standard deviation over fifteen random seeds

Relation	**Recall**	**Precision**	**F_1**
Causes	0,71	0,81	0,76
Detected_on	0,30	0,59	0,40
Dispersed_by	0,65	0,32	0,43
Affects	0,83	0,69	0,76
Found_on	0,72	0,68	0,70
Located_in	0,70	0,61	0,65
Transmits	0,75	0,43	0,55
ALL (micro)	0,70	0,59	0,64
ALL (macro)	0,66	0,59	0,61

Table 5. Relation extraction scores on the EPOP corpus using ReBERT with gold-standard entities.

The results indicate that relation extraction remains challenging. Performance is notably lower for relation types with few positive examples, such as Transmits and Dispersed_by, which each have fewer than 100 instances in the training set. For Detected_on and Located_in relations, despite having a comparable number of positive examples to other relations, performance remains limited. This may be due to their broader semantic scope, which leads to a larger and more heterogeneous set of negative examples thereby increasing classification difficulty.

We evaluated the performance of four LLMs on the DocRE tasks using EPOP training and development sets. The test set was deliberately excluded to preserve its integrity for future benchmarking. This prevents mitigates the risk of data leakage, as publicly released test data could be incorporated into future model training, biasing the results in subsequent evaluations.

Temperature and *top-p* are critical hyperparameters that influence the diversity and stability of model outputs. To ensure reliable results, we experimented with different values on a subset of the EPOP dataset to identify settings that maximize correct and valid answers. As expected, increasing temperature consistently reduced model consistency across repetitions and slightly decreased accuracy. Similarly, top-p values above 0.5 led to a further drop in accuracy. Based on these findings, we fixed the temperature at 0.2 and top-p at 0.1 in all experiments. Table 6 displays the scores obtained for the four models averaged over 5 runs per document, using strict string-matching for entity evaluation (see (Yao et al., 2026) for a detailed description of the evaluation protocol and output consistency analysis).

	GPT-4o-mini			Kimi			DeepSeek-V3			Qwen3.0		
	P	**R**	**F_1**	**P**	**R**	**F_1**	**P**	**R**	**F_1**	**P**	**R**	**F_1**
Affects	0.75± 0.41	0.50 ±0.37	0.57± 0.36	0.70±0.39	0.50± 0.37	0.56 ±0.36	0.82± 0.34	0.61± 0.35	0.68± 0.33	0.84± 0.27	0.64± 0.29	0.70± 0.26
Causes	0.71± 0.43	0.63 ±0.42	0.65± 0.42	0.78±0.37	0.68± 0.38	0.71 ±0.37	0.80± 0.39	0.69± 0.40	0.72± 0.39	0.76± 0.36	0.68± 0.37	0.70± 0.36
Has been found on	0.74± 0.32	0.75 ±0.39	0.60± 0.39	0.64±0.38	0.75± 0.37	0.59 ±0.38	0.84± 0.31	0.69± 0.34	0.74± 0.32	0.83± 0.29	0.70± 0.34	0.74± 0.32
Located in	0.81± 0.36	0.44 ±0.31	0.53± 0.31	0.88±0.30	0.51± 0.32	0.61 ±0.31	0.93± 0.22	0.53± 0.29	0.64± 0.26	0.91± 0.21	0.55± 0.28	0.64± 0.25
Transmits	0.86± 0.32	0.70 ±0.35	0.74± 0.33	0.82±0.27	0.66± 0.35	0.70 ±0.31	0.98± 0.09	0.79± 0.31	0.84± 0.23	0.87± 0.29	0.70± 0.36	0.75± 0.32
All (Micro)	0.55 ± 0.30	0.62 ± 0.30	0.54± 0.27	0.60± 0.30	0.61± 0.31	0.56 ± 0.27	0.64± 0.28	0.67± 0.27	0.61± 0.24	0.65± 0.26	0.52± 0.31	0.53± 0.25
All (Macro)	0.78± 0.05	0.57 ±0.09	0.63± 0.07	0.79±0.06	0.59± 0.07	0.64 ±0.06	0.87± 0.07	0.66± 0.09	0.72± 0.07	0.84± 0.05	0.65± 0.06	0.71± 0.04

Table 6: Scores in relation extraction in the EPOP corpus using LLM. P denotes Precision, R denotes Recall and F_1 is the harmonic mean of the two.

The four models achieve comparable scores, with Deep-Seek showing a 5-point advantage. Precision and recall are balanced despite the length of some texts, which could have affected recall. The F_1 score of 60 obtained with our simple one-shot hard-prompting strategy is very encouraging for a task of this level of difficulty, while also indicating room for improvement.

The proportion of malformed JSON outputs ranges from 6% to 10%, but almost all were repairable, except for about 1–2% in the case of Kimi and Qwen3. This is a positive outcome given the length and complexity of the response structure.

Some of the false positives are predictions that do not adhere to argument type constraints. For example, the arguments of *Found_on* must be of type *Species*. However, the models sometimes use a *Disease* type as the source argument, instead of using the correct relation *Expressed_by*. This type of confusion should be easy to correct through post-processing.

Despite explicit instructions to limit inference to the content of the text, some predictions exceed the information strictly conveyed within the document. As the boundary between implicitly evident content and external background knowledge remains ambiguous, such outputs are challenging to evaluate in the absence of a gold standard.

5. Discussion

The experiments conducted with the two approaches, BERT-based models and LLMs, were deliberately diversified to cover different evaluation and application scenarios. In the case of small models, relation predictions were made using gold-standard entities rather than predicted ones, in order to provide a modular comparison basis for relation extraction methods. For LLMs, the tasks were evaluated instead under real-world conditions, where relations are predicted from the entities generated by the model itself.

Ambiguity in the surface realization of semantic roles is a key challenge, particularly for disease/causative-agent and pest/vector pairs. LLMs outperform smaller models in resolving these distinctions.

While scores for named entity recognition and relation extraction are reported separately for small models due to the setting, it is important to

note that their combined end-to-end performance would be much lower due to NER error propagation. F_1 performance of LLMs in these experiments would then be significantly higher ; however, their computational cost is vastly greater. In a crop health monitoring context, where a few thousand documents are processed each week, the use of LLMs remains feasible. However, unlike small models, LLMs do not easily provide exact positions of extracted information in the text, which would require additional predictive post-processing to locate entities within the source text.

6. Conclusion and Future Work

We introduced the EPOP corpus, a new language resource for advanced information extraction in the under-resourced domain of plant health monitoring. EPOP addresses key challenges such as ambiguous entity mentions, normalization to large taxonomic and geographic references, and long-distance event extraction. Its structured annotation schema and rigorous design make it a high-quality benchmark for evaluating both traditional and large language models. The diversity of experiments conducted across tasks and state-of-the-art models provides valuable baselines and opens avenues for future research. EPOP also supports community evaluation through the PestCLEF task at the LifeCLEF Lab, CLEF 2026.

7. Limitations

The current version of the EPOP corpus presents limitations. First, the document selection covers a relatively short time span and focuses on 15 high-priority monitored species in Europe. This introduces a geographical, thematic and temporal bias that may limit generalizability. Future expansions should aim to increase both temporal coverage and taxonomic diversity to better reflect the breadth of plant health monitoring.

Second, the documents were translated in English from 25 source languages by GoogleTranslate. While necessary for the end-users, the potential impact of translation on information quality and extraction accuracy remains unassessed. In particular we observed that some infrequent disease and species vernacular names were not accurately translated. This introduces a potential source of bias, which may disproportionately affect the extraction of species and biotic interactions from regions where source languages are less well-supported by current machine translation systems.

Third, identity coreference sets are limited to spans of five consecutive sentences including titles. This constraint was chosen to balance annotation effort and practical scope, but it may lead to underestimation of LLMs performances that tend to merge semantically equivalent mentions beyond the annotated window in the DocRE evaluation.

The tests conducted with LLMs for entity normalization with large reference databases (NCBI Taxonomy and GeoNames) were inconclusive. A more thorough investigation of appropriate strategies for large-scale normalization with LLMs remains necessary.

These limitations reflect design choices made to enable timely corpus release and resource usability. However, they also identify clear directions for corpus extension and more robust evaluation in future work.

8. Availability statement

The training and development annotations of EPOP are publicly available in the BioNLP-ST format at https://doi.org/10.57745/ZDNOGF under CC-BY license.

EPOP document dataset is publicly available at https://doi.org/10.57745/YKSEPY

The code and prompts are available at https://github.com/YaoXinZhi/EPOP-Benchmark-IREC-2026

9. Acknowledgments

We thank the annotators for their careful and consistent work in labeling the EPOP corpus. Their contribution was essential to the development and validation of the dataset. Their deep domain and methodological expertise ensured the reliability and quality of the annotations. The complete list of contributors is available at doi:10.57745/ZDNOGF.

The authors acknowledge funding support from the French National Research Agency (ANR) for the BEYOND project (contract n° 20-PCPA-0002) and the EcoControl project (contract n° ANR-24-PEAE-0004 as well as the Chinese Scholarship Council and the International Mobility program of the DataIA Institute for the scholarships awarded to Xinzhi Yao.

This work was granted access to the HPC resources of Saclay-IA of Paris-Saclay University through the Lab-IA machine.

10. Bibliographical References

Bada, M., Eckert, M., Evans, D., Garcia, K., Shipley, K., Sitnikov, D., Baumgartner Jr., W. A., Cohen, K. B., Verspoor, K., Blake, J. A. and Hunter, L. E. (2012). Concept annotation in the CRAFT Corpus. *BMC Bioinformatics*, 13:161. doi:10.1186/1471-2105-13-161

Bossy, R., Jourde, J., Manine, AP. Veber P., Alphonse E., van de Guchte M., Bessières P., Nédellec C. (2012). BioNLP Shared Task - The Bacteria Track. *BMC Bioinformatics* 13 (Suppl 11), S3. doi : 10.1186/1471-2105-13-S11-S3

Bossy, R., Deléger, L., Chaix, E., Ba, M. and Nédellec, C. (2019). Bacteria Biotope at BioNLP Open Shared Tasks 2019, *Proceedings of the 5th Workshop on BioNLP Open Shared Tasks BioNLP-OST@EMNLP-IJNCLP 2019*, Hong-Kong, Association for Computational Linguistics. 10.18653/v1/D19-5719

Chaix, E., Dubreucq, B., Fatihi, A., Valsamou, D., Bossy, R., Ba, M., Zweigenbaum P., Bessières P., Lepiniec L., and Nédellec, C. (2016). Overview of the regulatory network of plant seed development (SeeDev) task at the BioNLP shared task 2016. In *Proceedings of the 4th BioNLP Shared Task Workshop*, Berlin, Germany. Association for Computational Linguistics. doi : 10.18653/v1/W16-3001

Kim, J.-D., Nédellec, C., Bossy, R, Deléger, L. (eds.). *Proceedings of the 5th Workshop on BioNLP Open Shared Tasks (BioNLP-OST)@EMNLP-IJCNLP 2019*, Hong Kong, China, November 4, 2019. Association for Computational Linguistics. https://aclanthology.org/D19-5700/

Kim M., Chae K., Lee S, Jang H. J. and Kim S. (2020). Automated Classification of Online Sources for Infectious Disease Occurrences Using Machine-Learning-Based Natural Language Processing Approaches. *Int J Environ Res Public Health*. 17(24):9467. doi: 10.3390/ijerph17249467

Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C. H., & Kang, J. (2020). BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics (Oxford, England)*, *36*(4), 1234-1240.

Lotreck, S., Segura Abá, K., Lehti-Shiu, M. D., Seeger, A., Brown, B. N., Ranaweera, T., Schumacher a., Ghassemi M., Shiu, S. H. (2024). Plant Science Knowledge Graph Corpus: a gold standard entity and relation corpus for the molecular plant sciences. *in silico Plants*, *6*(1) doi : 10.1093/insilicoplants/diad021

Luoma, J., Nastou, K., Ohta, T., Toivonen, H., Pafilis, E., Juhl Jensen, L. and Pyysalo, S. (2023) S1000: a better taxonomic name corpus for biomedical information extraction. *Bioinformatics* 39(6). doi:10.1093/bioinformatics/btad369

Miranda-Escalada, A., Mehryary, F., Luoma J., Estrada-Zavala, D., Gasco, L., Pyysalo, P., Valencia, A. and Krallinger, M., (2023). Overview of DrugProt task at BioCreative VII: data and methods for large-scale text mining and knowledge graph generation of heterogenous chemical–protein relations, *Database*, Volume 2023. doi : 10.1093/database/baad080

Nainia, A., Vignes-Lebbe, R., Chenin, E., Sahraoui, M., Mousannif, H. and Zahir, J. (2024). FloraNER: A new dataset for species and morphological terms named entity recognition in French botanical text. *Data in Brief*, *56*. doi: 10.5281/zenodo.10940912.

Nédellec, C., Sauvion, C., Bossy, R., Borovikova, M. and Deléger, L. (2024). TaeC: A manually annotated text dataset for trait and phenotype extraction and entity linking in wheat breeding literature. *Plos one*, *19*(6). doi : 1371/journal.pone.0305475.

Newton, G., Korol, O., Lévesque, A., Favrin, R. and Graefenham, T. (2019). Extracting pest risk information from risk assessment documents. In *2019 ACM/IEEE Joint Conference on Digital Libraries (JCDL)* (pp. 368-369). IEEE. doi: 10.1109/JCDL.2019.00074

Nguyen, N. T., Gabud, R. S. and Ananiadou, S. (2019). COPIOUS: A gold standard corpus of named entities towards extracting species occurrence from biodiversity literature. *Biodiversity data journal*, (7). doi : 10.3897/bdj.7.e29626

Papazian F., Bossy R. and Nédellec C., « AlvisAE: a collaborative Web text annotation editor for knowledge acquisition », *The 6th Linguistic Annotation Workshop (The LAW VI),* p 149-152, Jeju, Corée, 2012. http://www.aclweb.org/anthology/W12-3621

Roche, M., Rabatel, J., Trevennec, C., and Pieretti, I. (2024). PADI-web for Plant Health Surveillance. In *International Conference on Advanced Information Systems Engineering* (pp. 148-156). Cham: Springer Nature Switzerland. doi: 10.1007/978-3-031-61000-4_17

Shankar, P., Bitter, C., & Liwicki, M. (2020). Digital crop health monitoring by analyzing social media streams. In *2020 IEEE/ITU International Conference on Artificial Intelligence for Good (AI4G)* (pp. 87-94). IEEE. doi: 10.1109/AI4G50087.2020.9310985

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

Tang, A. (2023). *Leveraging linguistic and semantic information for relation extraction from domain-specific texts.* Doctoral dissertation, INRAE, Université Paris-Saclay. https://theses.hal.science/tel-04420517

Yao X., Nédellec C., Xia J. and Bossy R. (2026). Consistency–accuracy correlation in hard-prompted LLMs for entity and relation extraction: empirical findings from plant-health data. *Genom. Inform.* 24, 3. doi: 10.1186/s44342-025-00063-2

Yuan, W., Yang, W., He, L., Zhang, T., Hao, Y., Lu, J. and Yan, W. (2024). Research on Entity and Relationship Extraction with Small Training Samples for Cotton Pests and Diseases. *Agriculture*, 14(3), 457. doi: 10.3390/agriculture14030457

Zhang, D., Zheng, G., Liu, H., Ma, X. and Xi, L. (2023). AWdpCNER: Automated Wdp Chinese Named Entity Recognition from Wheat Diseases and Pests Text. *Agriculture*, *13*(6), 1220. doi: 10.3390/agriculture13061220

Zhang, J., Guo, M., Geng, Y., Li, M., Zhang, Y., and Geng, N. (2021). Chinese named entity recognition for apple diseases and pests based on character augmentation. *Computers and Electronics in Agriculture*, *190*, 106464. doi : 10.1016/j.compag.2021.106464

Zhang, L., Nie, X., Zhang, M., Gu, M., Geissen, V., Ritsema, C. J., Dangdang N. and Zhang, H. (2022). Lexicon and attention-based named entity recognition for kiwifruit diseases and pests: A Deep learning approach. *Frontiers in plant science*, *13*, 1053449. doi: 10.3389/fpls.2022.1053449

Zhong, Z., and Chen, D. (2021). A Frustratingly Easy Approach for Entity and Relation Extraction. In *2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021* (pp. 50-61). Association for Computational Linguistics (ACL). doi: 10.18653/v1/2021.naacl-main.5

11. Language Resource References

Barbaresi, A., *Trafilatura*. (2021) Apache 2.0 license. https://trafilatura.readthedocs.io/en/latest/

R. Bossy, S. Duperier, I. Pieretti, L. Deléger. M. Grosdidier, C. Nédellec, (2025) *EPOP Training and development datasets for information extraction in plant epidemiomonitoring*" doi: 10.57745/ZDNOGF, Recherche Data Gouv.

R. Bossy, C. Nédellec, L. Deléger, M. Courtin. Knowledge Graph Extraction on Plant Pests from Web Documents https://www.imageclef.org/PestCLEF2026

C. Nédellec, G. Catalina, L. Elisa, Grosdidier M., L.Deléger, R. Bossy, S. Duperier, C. Sauvion, I. Pieretti. *Guidelines for the annotation of the corpus Epidemiomonitoring Of Plant (EPOP).* 2024. https://hal.inrae.fr/hal-04744299

C. Nédellec, M. Grosdidier, S. Duperier, I. Pieretti, R. Bossy, L. Deléger, C. Sauvion. (2025). *EPOP Epidemiomonitoring of plant dataset corpus*. doi: 10.57745/YKSEPY, Recherche Data Gouv.