



HAL
open science

Guidelines for the annotation of the corpus Epidemiomonitoring Of Plant (EPOP)

Claire Nédellec, Garcia Catalina, Lubrini Elisa, Grosdidier Marie, Louise Deléger, Robert Bossy, Sandy Dupérier, Clara Sauvion, Isabelle Pieretti

► To cite this version:

Claire Nédellec, Garcia Catalina, Lubrini Elisa, Grosdidier Marie, Louise Deléger, et al.. Guidelines for the annotation of the corpus Epidemiomonitoring Of Plant (EPOP). 2024. hal-04744299

HAL Id: hal-04744299

<https://hal.science/hal-04744299v1>

Submitted on 18 Oct 2024

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

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



Distributed under a Creative Commons Attribution - ShareAlike 4.0 International License

Guidelines for the annotation of the corpus *Epidemiomonitoring Of Plant* EPOP

Version - 1.3

18 October 2024

Authors

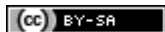
Claire Nédellec, Catalina Garcia, Elisa Lubrini, Marie Grosdidier, Louise Deléger, Robert Bossy, Sandy Duperier, Clara Sauvion, Isabelle Pieretti.

For

Expert annotators involved in the annotation campaign

Users of the annotated corpus, information extraction training and evaluation

Copyright 2024 by INRAE.



The documents Annotation guidelines for the EPOP corpus is distributed under the terms of the Creative Attribution-ShareAlike 4.0 License (CC-BY-SA). <http://creativecommons.org/licenses/by-sa/4.0/>

Content

1.	Introduction.....	3
2.	Annotation of entities.....	3
2.1.	Entity types	3
2.2.	General guidelines for entity annotation.....	3
2.2.A.	Boundaries	3
2.2.B.	Enumeration	4
2.2.C.	Discontinuity.....	4
2.2.D.	Acronyms	5
2.2.E.	Quotes	5
2.2.F.	Annotation zone	6
2.3.	Annotation guidelines by entity type	6
2.3.A.	Species boundaries.....	6
2.3.B.	Pest.....	7
2.3.C.	Host plant.....	8
2.3.D.	Vector.....	11
2.3.E.	Disease	12
2.3.F.	Dissemination pathway.....	14
2.3.G.	Location	15
2.3.H.	Date.....	18
3.	Normalisation.....	19
3.1.	Normalisation of organisms.....	19
3.2.	Normalisation of dissemination pathways.....	21
3.3.	Normalisation of location.....	22
4.	Annotation of relationships.....	25
4.1.	Coreference	25
4.2.	Thematic binary relationship	26
4.2.A.	Causes	27
4.2.B.	Found_on	27
4.2.C.	Vected_by	28
4.2.D.	Expressed_by	28
4.2.E.	Dispersed_by.....	28
4.2.F.	Located_in.....	29
4.2.G.	Detected_by	29
4.3.	N-ary relationships (events).....	30

4.4.	Modality	30
4.4.A.	Negation	30
4.4.B.	Hypothesis.....	31

1. Introduction

This document details the instructions for manual annotation of the EPOP (*Epidemiomonitoring Of Plant*) corpus of documents. It is intended for the experts who annotate these documents and for the evaluation and development of methods that automatically predict these annotations. It is divided into three parts, dedicated to entity annotation and normalization, and relationship annotation. Each instruction is detailed and illustrated with examples.

2. Annotation of entities

2.1. Entity types

Pest	Dissemination_path way	Location
Vector	Disease	Date
Plant		Quantity

In the rest of the document, we use this color code in the examples to highlight entity mentions and refer to annotated entities.

2.2. General guidelines for entity annotation

2.2.A. Boundaries

We call the boundary the limit of the text span. Annotations should extend to all relevant elements but exclude those that do not provide information to identify the entity. In particular, words that are of no interest out of context are not annotated. Annotations are restricted to a nominal group. They may include a relative proposition if it contributes to the description of the entity.

Detailed instructions are given below in the sections dedicated to the different types. This section presents the general principles, illustrated by a few examples.

Examples

	Correct	Incorrect
1. <i>Distinct entities</i>	for the control of HLB; in the genetic improvement of citrus varieties with resistance to <i>Candidatus Liberibacter asiaticus</i>	or the control of HLB; in the genetic improvement of citrus varieties with resistance to <i>Candidatus Liberibacter asiaticus</i>

- | | | | |
|----|-------------------------------|--------------------------------------|--------------------------------------|
| 2. | Type of the geographic entity | Basque Country | Basque Country |
| 3. | Type | Huanglongbing disease | Huanglongbing disease |
| 4. | description (not annotated) | found on the steep Mont Blanc massif | found on the steep Mont Blanc massif |
-

2.2.B. Enumeration

If the items enumerated are all independently relevant, they are annotated separately (example (5)).

If an enumeration of terms qualifies the same entity, the terms are annotated separately (example (6)). If the enumeration is made up of increasingly specific or general terms, only the most specific entity participates in the relations.

Examples

		Correct	Incorrect
5.	Distinct entities	The bacteria can be found in both Italy and France	The bacteria can be found in both Italy and France
6.	Region, country	During the 2014 - 2015 harvest, field samples of wilting chickpea plants were taken to the laboratory at the CENSA, Mayabeque Province, Cuba.	During the 2014 - 2015 harvest, field samples of wilting chickpea plants were taken to the laboratory at the CENSA, Mayabeque Province, Cuba.

2.2.C. Discontinuity

Entity annotations can be discontinuous, especially when the same adjective or noun qualifies two distinct entities. In example (7) two distinct regions of Tuscany are mentioned. Discontinuity will not be used when the entity is represented by the union of words or phrases, as in example (8). This method should not be used to arbitrarily bring expressions in the text closer together, as in example (9), where the intention might be to exclude 'fungus called'. Here, only 'tropical race 1' should be annotated.

Examples

		Correct	Incorrect
7.	2 discontinuous entities	North-Western provinces and Southern provinces of Tuscany	North-Western provinces and Southern provinces of Tuscany

8.	Duration, a single entity	between August 15 and September 12	between August 15 and September 12
9.	No discontinuity	a strain of Fusarium fungus called tropical race 1	a strain of Fusarium fungus called tropical race 1

Exceptions

If the relevant text span contains non-meaningful characters, such as tags, the entity annotation includes the characters in question, and is not discontinuous (example (10)).

Example

10. infected by different subspecies of `Xylella` (`multiplex`

2.2.D. Acronyms

Acronyms that follow an entity reference and refer to the same entity are annotated separately. The two entities are connected by a coreference relationship, see the Coreference section 4.1.

Examples

11. `Fusarium oxysporum f. sp. Cubense` fungi (`Foc`)
12. `Huanglongbing` (`HLB`)

2.2.E. Quotes

Any quotation marks or apostrophes around the entity are not included in the annotation. Conversely, if they are inside the entity, they are included in the annotation according to the discontinuity rule above.

	Correct	Incorrect
13.	External double quotes <code>"Bois noir"</code>	<code>"Bois noir"</code>
14.	External quotes <code>'Candidatus Phytoplasma solani'</code>	<code>'Candidatus Phytoplasma solani'</code>

2.2.F. Annotation zone

All text in the document is annotated by default. However, if useful, a part of the text can be excluded for reasons such as the text is too long, repetitive, irrelevant, etc. The relevant part of the text is tagged by "annotated text".

Entities within the reference title may be linked together, but there must be no relationship linking entities in the document text with entities in the references, nor between entities in separate references.

2.3. Annotation guidelines by entity type

2.3.A. Species boundaries

When qualifying an entity that is not a scientific name, entity type names (insect, pest, bacteria, ...) and relevant adjectives are included in the annotation to help specify the entity (examples (16) to (19)). When entities are specified by scientific taxon names, the scientific name is sufficient to designate the taxon, so entities are annotated excluding entity type and qualifiers (examples (20), (21) and (22)).

Examples

		Correct	Incorrect
16.	Irrelevant adjective	Field samples of wilting chickpea plants with chlorotic and necrotic roots were taken to the laboratory at CENSA.	Field samples of wilting chickpea plants with chlorotic and necrotic roots were taken to the laboratory at CENSA.
17.	include relevant information for the specification	PWD occurs when the host plant is infected with pinewood nematodes.	FPWD occurs when the host plant is infected with pinewood nematodes.
18.	Full name	yellow mottle virus	yellow mottle virus
19.	type	fall armyworm pest	fall armyworm pest
20.	type	<i>Bactrocera dorsalis</i> insect	<i>Bactrocera dorsalis</i> insect
21.	type	<i>Xylella fastidiosa</i> bacterium	<i>Xylella fastidiosa</i> bacterium
22.	Irrelevant adjectives	Gram-negative bacterium <i>Candidatus Liberibacter</i> species	Gram-negative bacterium <i>Candidatus Liberibacter</i> species

Note the difference between adjectives that add information to specify the entity (e.g. *fall armyworm pest*) and adjectives that describe the entity without giving more information about its identity (e.g. (1) *wilting chickpea*).

2.3.B. Pest

Definition

A pest is a living organism or group of organisms whose harmfulness is explicitly stated in the document. These are animals or pathogens, of the genera bacteria, viruses, insects, fungi, oomycetes, nematodes, gastropods or arachnids. More rarely, a pest may also be a parasitic plant.

They are referred to by their full scientific or vernacular (common) name or as a metaphor, acronym or abbreviation.

References to non-harmful organisms should not be annotated, and genera that include non-pathogenic species should also not be annotated (e.g., *Fusarium* sp.). Genera are annotated only if all genus species are pest.

When the entity designates a pest by its taxon, only entries for taxa equal to or below the taxonomic level of the genus are annotated, i.e. genera, species and subspecies (e.g. strains, cultivars, pathovars). Taxa above the genus level are not annotated.

Annotated pest entities must all be normalized (see the [normalisation section](#)).

Examples

- (23) **Scientific name** The *Spodoptera frugiperda* is a species in the order Lepidoptera.
- (24) **Vernacular name** The name “*fall armyworm*” can refer to several species.
- (25) **Vernacular name (acronym)** An urgent response to the rapid spread of **FAW** is needed.
-

Boundaries

The pest annotation includes all qualifiers useful in determining the organism and excludes all others. In particular, the scientific name includes words designating the subspecies, the authority and the date where applicable.

Examples

- | | Correct | Incorrect |
|------------------------|--|--|
| (26) Subspecies | <i>X. fastidiosa subspecies pauca</i> ST53 | <i>X. fastidiosa</i> subspecies pauca ST53 |
| (27) Authority | <i>Spodoptera frugiperda</i> (JE Smith) | <i>Spodoptera frugiperda</i> (JE Smith) |

(28)

Date [Spodoptera frugiperda, Smith, 1797](#)

[Spodoptera frugiperda, Smith, 1797](#)

Pests designated by their stage of development are not annotated.

Examples

		Correct	Incorrect
(29)	Development stage	The larvae develop in 10 months	The larvae develop in 10 months
(30)	information not needed to determine the species	adult Popillia japonica	adult Popillia japonica

Exceptions

Some pest descriptions are very general and do not provide any information at all. The following statements are in this category. They are not annotated if they are used alone, without any other precision, i.e. without any complement or adjective.

- | | | |
|-------------|-----------------|------------|
| - pest | - microbial | - snail |
| - specimens | - microorganism | - nematode |
| - biotope | - population | - pathogen |
| - carrier | - subject | - parasit |
| - host | - vector | |
| - microbe | - virus | |

2.3.C. Host plant

Definition

Plant entities are host plants, a plant or group of plants whose role as host is explicit in the document. The host role may be expressed by a relationship with a pest or by terms characterizing its role.

The host role is to be understood here in the broad sense of a host organism that is the habitat of a pest and includes resting plants, reservoir plants, and plants in which the pest reproduces. Annotated entities must be normalized (see the [normalisation section](#)).

Examples

(31) [...] shared by two hosts, [Crepis foetida](#) and [Daucus carota](#).

Species and genera can be denoted by their scientific name, their vernacular name or by the word that designates the whole of the same cultivated species (example (33)).

Examples

- (32) Pitch pine (*Pinus rigida* Mill.)
- (33) around the vineyards in the Tuscan region.
-

The host plant type does not include plants designated by their use or industry, e.g. ornamental, cultivated, garden plant, etc. The host plant type does not include plant parts, e.g. fruits, or stems.

Plant names in disease names are not annotated. For example, in "Tomato brown rot virus", "Tomato" should not be annotated as a plant.

Counter-examples

- (34) common plant species such as olive, blackberry, plum, avocado, citrus, cypress and many other woody species". "as major pest of many fruit crops", "the oriental fruit fly is known to target over 230 different types of products, including pome, stone fruits, citrus, dates, avocados, peppers and tomatoes
- (35) a fungus called TR1 began destroying banana crops.
-

When the entity refers to a plant by its taxon, only entities less than or equal to the genus are annotated, i.e. genera, species and subspecies (e.g. cultivar). Genotypes in parentheses are not annotated (ex 37). Genetically modified organisms (GMOs) are annotated as plants, including their names (ex 38).

Examples

	Correct	Incorrect
(36) cultivar	Asplenium sp. CV 961 NCBI ID: 218655 (Asplenium sp. CV 961)	Asplenium NCBI ID: 32071 (Asplenium)
(37) cultivar and genotype	banana cultivar Maça (Silk, AAB)	banana cultivar Maça (Silk, AAB)
(38) GMO	Bt maize has been granted environmental release in Kenya. TELA maize has the potential to transform Nigeria	Bt maize has been granted environmental release in Kenya. TELA maize has the potential to transform Nigeria

Boundaries

Mention of the host plant type includes all qualifiers useful for its determination and excludes others, as well as qualifiers indicating that it is a cultivated species (crops, farms, plantation, etc.) or a variety. In particular, the scientific name includes words designating the subspecies, authority and date where applicable.

Examples

- (39) *Plectranthus hadiensis* var. *tomentosus* plants with stem and root rot symptoms
- (40) A sentinel plantation of *Prunus domestica* cv. *Opal*, *Quercus petraea* and *Salix alba* were established .
- (41) *Pinus koraiensis* Siebold & Zucc
- (42) it is still largely unknown which *African banana varieties* are susceptible
- (43) in addition, certain *varieties of plane trees* have been selected
- (44) *banana variety called Cavendish*
-

Exceptions

Plant names included in disease names are not annotated. For example, in "Tomato brown rot virus", "Tomato" should not be annotated as a plant.

Some terms designate very general entities and provide no information whatsoever. The following terms are not annotated if they are used alone, without any further clarification, i.e. without any complement or adjective:

- host
- crop

In some cases, the vernacular name of the plant is used to characterize the sector, the industry, etc. In this case, the plant name will not be annotated. However, plants associated with a farm, a farmer or a plantation are annotated.

Examples

	Correct	Incorrect
(45) fruit	The global supply of bananas has previously been threatened	The global supply of <i>bananas</i> has previously been threatened

(46)	fruit	Bananas are Britain's favourite fruit	Bananas are Britain's favourite fruit
(47)	fruit	They often grow the Cavendish banana that is sold in supermarkets	They often grow the Cavendish banana that is sold in supermarkets
(48)	workers	The Fusarium fungus is well-known within the world of banana growers	The Fusarium fungus is well-known within the world of banana growers
(49)	farmers	where the dominant TR4 strain is affecting banana farmers	where the dominant TR4 strain is affecting banana farmers
(50)	plantations	save banana plantation with a well-known fungicide	save banana plantations with a well-known fungicide
(51)	farms	This pathogen affects tomato farms	This pathogen affects tomato farms

2.3.D. Vector

Definition

A vector is an organism, or a group of organisms, whose role as vector or carrier of a pest or disease to a plant is explicit in the document. This role may be represented by a relationship with a pest or plant, or a disease, or by terms characterizing its role, such as the word *vector*. Vectors can be insects, arachnids, nematodes, fungi and potentially other organisms. The role of the vector is to be understood here in a broad sense, including external (e.g. thorax hair) or internal (saliva, mouth parts) transport. The transmission capacity of a pest to a given plant is not necessarily mentioned. When the entity refers to the vector by its taxon, only taxa less than or equal to the genus are annotated, i.e. genera, species and subspecies.

Annotated vector entities must be normalized (see the [normalisation section](#)).

Examples

-
- (52) ACP vectors the lethal bacterium (*Candidatus Liberibacter asiaticus*) that causes Huanglongbing
- (53) with respect to the Citrus Leafhopper plague (*Diaphorina citri*. kuwayama), for which prevention
- (54) The African citrus psyllid, *Trioza erytreae* (Hemiptera: Triozidae) is a vector of citrus greening disease (Huanglongbing - HLB) caused by the bacterium *Candidatus liberibacter*
-

Vector species and genera are denoted by their scientific name, vernacular name, acronym, abbreviation, or metaphor.

Examples

- (55) Asian citrus psyllid
- (56) *T. erytrae*, *Diaphorina citri*
-

Boundaries

The vector type includes all qualifiers useful in determining the species but excludes others. In particular, the scientific name includes the words designating the subspecies, the authority and the date if applicable. In the example (57), the references (Coleoptera: Cerambycidae) in brackets after (Gebler) are the higher taxonomic ranks, not annotated.

Examples

- (57) *Monochamus alternatus* Hope and *M. saltuarius* (Gebler) (Coleoptera : Cerambycidae) transmit *Bursaphelenchus xylophilus* causing pine wilt disease
-

Exceptions

Some entities designate very generic species and provide no information. The following terms are not annotated if they are used on their own, without any further specification, i.e. without any complement or adjective:

- vector
- insect
- insect vector

2.3.E. Disease

Definition

A disease can be the consequence of the presence of a harmful organism on, or in, a host plant. It is designated by a specific term. The Disease entity type excludes symptoms or signs when alone.

Examples

- (58) Huanglongbing, citrus greening disease, pine wilt disease, beech bark disease, Olive Quick Decline Syndrome (OQDS), almond leaf scorch disease (ALSD)
- (59) Common Florida Citrus Disease

Terms designating a plant or part of a plant, followed by the word "disease", are not annotated if they do not represent the usual name given to the disease, e.g. *apple tree disease* should not be annotated.

In some cases, however, these expressions refer to specific, well-identified diseases, and should be annotated. For example,

- *banana disease* usually means *Panama disease*

Boundaries

The mention of disease is restricted to the words that designate it. Qualifiers representing the severity or extent of the disease are not included in the annotation. The plant name is included in the disease annotation if it is part of the disease name.

Examples

		Correct	Incorrect
(60)	severity	deadly pine wilt disease	deadly pine wilt disease
(61)	extent	widespread pine wilt disease	widespread pine wilt disease
(62)	The name of the plant is part of the disease name	pine wilt disease	pine wilt disease

Exceptions

Disease names in the organism names are not annotated. For example, in the sentence "*tomato brown rot virus has been found in a tomato farm*", "*tomato brown rot virus*" is annotated as a pest. The disease "*tomato brown rot*" should not be annotated.

Names of groups or families of diseases are not annotated, if they refer to several diseases sharing a certain symptom or affecting a certain plant species. For example: "*yellowing diseases*", "*vine disease*" or "*grapevine yellows*".

Some terms designate a very generic entity. These mentions are so frequent that they do not provide any information. The following terms are not annotated if they are used alone, without any other precision, i.e. without any complement or adjective:

- disease
- illness

2.3.F. Dissemination pathway

Definition

Dissemination pathways are either abiotic or biotic pathways. Their role is made explicit in the document. It may be represented by a relationship with the pest or disease transmitted or with the host plant, or by terms characterizing its role, such as the word *pathway*. These are generally parts of plants (e.g. seeds, fruits), whether processed or not (e.g. wooden transport pallets), or more broadly abiotic pathways: air, sea, water, roads, etc.

The dispersal route by which a pest affects a plant is not necessarily mentioned in the document.

Vectors are living organisms; they are not annotated as dissemination pathways. Farming practices are not annotated in general, but farming tools or equipment are annotated if the transmission through them is explicit.

Annotated pathway entities must be normalized (see the [normalisation section](#)).

Examples

- (63) Imports are vital to Florida's economy and can include hitchhikers within the **cargo** or on/in the vessel (e.g. **woodboring beetles** in **wooden pallets**, organisms in **ship ballast water**).
- (64) It is mainly transmitted mechanically by **contact between plants**, with cultivation care (tying, weeding, picking, etc.).
- (65) It is transmitted by **hands** that have been in contact with an infected plant, on **contaminated clothing**, on **tools** used on infected plants
-

Boundaries

Dissemination pathway entity span includes all qualifiers useful in determining its meaning, but excludes all others. Pathways are annotated with their associated movement terms when relevant.

Examples

- (66) The most common pathway for these pests to enter the State is by "hitchhiking" in **fruits** and **vegetables** brought back inadvertently by **travelers as they return**.
- (67) The main route of entry would be the importation of **fruits** from affected areas as well as **contaminated plant material** (**plants** or **twigs** for grafting not subject to quarantine)

Exceptions

Some mentions designate a very generic or abstract entity as the following terms, which are not annotated if they are used alone, without any further specification, i.e. without any complement or adjective which characterize the pathway:

- transport
- carrier
- pathway
- trade contamination patterns

Normalisation

Annotated vector entities must be normalized (see the [normalisation section](#)).

2.3.G. Location

Definition

Entities of the Location type designate geographical locations and should be annotated only if they reference elements of a geographical nature.

Location can be designated by adjectives, e.g. *Apulian* region or *Brazilian* agriculture (68). They are annotated as Location when they are part of another entity, provided that the Location mentioned is the physical place and not an abstract entity. Adjectives characterizing a crop or plant, indicating the geographical location of these crops, are annotated as Location (e.g. (69) and (70)). Person nationalities (e.g. "Belgian and Dutch farmers" (71) and (72)) are not annotated.

However, in the case of metonymy, the nationality adjective would be annotated as a location, as in the example (74) where "growers" actually refers to plants cultivated by farmers.

Location entities must be normalized (see the [normalisation section](#)).

Examples

	Correct	Incorrect
(68)	Australian vineyards	Australian vineyards
(69)	Brazilian agriculture	Brazilian agriculture
(70)	Mediterranean vegetation	Mediterranean vegetation
(71)	The Spanish citrus employers	The Spanish citrus employers
(72)	Belgian et dutch farmers	Belgian et dutch farmers

- | | | |
|------|--|--|
| (73) | the animal of East Asian origin | the animal of East Asian origin |
| (74) | "A total of 57 Dutch growers have been infected since the first outbreak" | A total of 57 Dutch growers have been infected since the first outbreak" |
-

Boundaries

The annotations include all the details needed to identify the location. See examples (75) to (82).

Examples

	Correct	Incorrect
(75)	Discontinuous has colonized African and Asian countries	has colonized African and Asian countries
(76)	Including countryside 38 infected plants, 16 are in the countryside of Polignano	38 infected plants, 16 are in the countryside of Polignano
(77)	Including Province advance of Xylella in the province of Bari	advance of Xylella in the province of Bari
(78)	Relevant precision was also detected near the French border, in Basel (Switzerland)	was also detected near the French border, in Basel (Switzerland)
(79)	Relevant precision is now advancing –as it already did further north in Portugal - inland, towards Huelva	is now advancing –as it already did further north in Portugal - inland, towards Huelva
(80)	Relevant precision this nematode is widespread in the Far East and Siberian regions of Europe and the Russian Federation .	this nematode is widespread in the Far East and Siberian regions of Europe and the Russian Federation .
(81)	Together The bacteria was found in an ancient administrative region that now encompasses territories across Italy and France	The bacteria was found in an ancient administrative region that now encompasses territories across Italy and France
(82)	Including city city of Zürich	city of Zürich

Exeptions

Expressions referring to institutional, political or economic entities are not annotated (example (83) to (85)).

Some names of organizations, journals or institutes include geographical place names (examples (86) to (89)), as well as certain diseases or species names (scientific or vernacular), (examples (90) to (92)). These geographical place names should not be annotated.

Examples

	Correct	Incorrect
(83)	The head of the Balearic Islands Agriculture Service, Andreu Juan, also participated in the reception for the Corsican delegation	The head of the Balearic Islands Agriculture Service, Andreu Juan, also participated in the reception for the Corsican delegation
(84)	China's Agriculture Minister	China's Agriculture Minister
(85)	commitments set forth in the European strategy (...) do not overlap with European production	commitments set forth in the European strategy (...) do not overlap with European production
(86)	A team of researchers from Wageningen University	A team of researchers from Wageningen University
(87)	organized by the Inter-American Institute for Cooperation on Agriculture (IICA)	organized by the Inter-American Institute for Cooperation on Agriculture (IICA)
(88)	The Florida Department of Agriculture and Consumer Services (FDACS)	The Florida Department of Agriculture and Consumer Services (FDACS)
(89)	the Union believes that ratifying or signing agreements by the EU, such as with Mercosur or South Africa	the Union believes that ratifying or signing agreements by the EU, such as with Mercosur or South Africa
(90)	Panama disease	Panama disease
(91)	efficiency of the control of the African Psila	efficiency of the control of the African Psila
(92)	also known as the giant African snail	also known as the giant African snail

Exceptions

The following expressions are too general to be useful and are not annotated:

- worldwide
- all Southern countries

The addresses of authors of scientific articles are not annotated as places. They are usually located just after the title.

Examples

Correct

Incorrect

- (93) The vector is found worldwide. The vector is found worldwide.
- (94) in many countries, such as France and Spain. in many countries, such as France and Spain.

2.3.H. Date

Definition

Date statements are temporal expressions that can be a whole or partial date (year-month-day or just year) or a period. They can be absolute, or relative.

Examples

		Correct	Incorrect
(95)	Observation date	In 1948, the pine wilt disease was found on 3 pine trees in the region.	Spodoptera frugiperda, Smith, 1797
(96)	Relative period	This last decade, the disease spread	

Boundaries

The annotation must include the most precise date (time, day, year, month/season, etc.), as well as any prepositions included in the temporal expressions, for example, “from”, “over”, “between”, “during”, “more than” “since”, “begins in”, “within”, “during”, “in”, “at”, “for”, “by”, “of”, “on”, etc. Les dates relatives or dependents are also annotated.

Examples

		Correct	Incorrect
(97)	Observation date	In March 1948, the pine wilt disease was found on 3 pine trees in the region.	In March 1948, the pine wilt disease was found on 3 pine trees in the region.
(98)	Relative date	In March 1948, the disease was discovered. The following year its presence was recorded in 5 different European countries.	In March 1948, the disease was discovered. The following year its presence was recorded in 5 different European countries.
(99)	Indefinite period	over several years	over several years
(100)	Discontinuous	During the 1940s and 1950s	During the 1940s and 1950s

(101)	Season	During the spring	During the spring
(102)	Day	The National Bank for Economic and Social Development (BNDES) launched this Tuesday (30) the second public notice	The National Bank for Economic and Social Development (BNDES) launched this Tuesday (30) the second public notice
(103)	Since a date	available from the beginning of 2024	available from the beginning of 2024

Time intervals are annotated as single entities.

Examples

	Correct	Incorrect
(104)	Period In the period January 2017 - March 2020 , 60480 news items on 494 pests were selected.	In the period January 2017 - March 2020 , 60480 news items on 494 pests were selected.
(105)	Period between August 15 and September 12	between August 15 and September 12

Exceptions

Dates mentioned in bibliographic citations (whether in the body of the text or in the reference section) and dates in species names are not annotated.

3. Normalisation

3.1. Normalisation of organisms

Organism-type entities (i.e. pests, vectors and plants) are associated with their most precise possible taxon, which is designated ("standardized") by its taxonomic identifier. The chosen repository is the NCBI taxonomy, whose browser is at: <https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi>

If the organism is not in the NCBI taxonomy, it must be annotated with the Encyclopedia of Life (EOL) identifier: <https://eol.org/>

If the organism is not in either repository, it is normalized by the OntoBiotope ontology class, designated by its identifier. The OntoBiotope version is October 2021 at <https://agroportal.lirmm.fr/ontologies/ONTOBIOTOPE>

Taxa designated by an abbreviation, acronym or ellipsis are annotated by the identifier of the referenced taxon.

Taxa normalization takes into account the context of the document, which requires precise interpretation by the expert. For example, a pest entity "Xylella" without its context could be normalized by the genus Xylella; nevertheless, it may represent the coreference of a more

precise mention, or *Xylella* may also be the vernacular name, the common name, of the bacterium, in which case *Xylella* is normalized as *Xylella fastidiosa*, or a more precise subspecies, depending on the sense of the context.

In example (106), *Xylella* refers to *Xylella fastidiosa* and should therefore be annotated with the NCBI identifier of the *Xylella fastidiosa* species.

Example

(106) In the event of an outbreak of *Xylella fastidiosa*, all plants sensitive to *Xylella* must be destroyed

When a document designates the organism by its vernacular name, and it represents several species of the same genus, the genus is used for normalization, unless the species is clearly mentioned in the document, or the expert considers it so.

In example (107), citrus tree entity is normalized with the *Citrus* sp. genus identifier, as this genus includes all cultivated and non-cultivated citrus species (oranges, lemons, clementines, etc.).

Species normalization applies in cases where taxa can be differentiated at the subspecies level. For example, the entity *vineyards* (example (108)), would be normalized with the identifier of *Vitis vinifera*, because the entity is a crop and cultivated vines are differentiated at cultivar level, so *Vitis vinifera* includes them all.

However, if the entity and the context are not sufficient to know whether the taxon refers to a cultivated or wild plant, normalization would be made with the taxa that includes both cultivated and non-cultivated plants. In example (109), olive trees without precision would be normalized with the identifier of *Olea europaea*, while the entity olive groves (cultivated) would be normalized with the identifier of the cultivated subspecies, *Olea europaea subsp. Europaea* (110).

Examples

-
- | | | |
|-------|---------------------|---|
| (107) | <i>Citrus trees</i> | <i>Citrus</i> sp. ID 2706 |
| (108) | <i>Vineyards</i> | <i>Vitis vinifera</i> ID 29760 |
| (109) | <i>Olive trees</i> | <i>Olea europaea</i> ID 4146 |
| (110) | <i>Olive groves</i> | <i>Olea europaea subsp. europaea</i> ID 1583383 |
-

If the author of the text confuses two species, this error should be normalized as it would have been correct. In example (111), *Fusarium oxysporum f.sp. cubense tropical race 4* is the full scientific name of the species, the document presents it as a synonym of "race four tropical" while NCBI taxonomy considers them as different. In this case, the two entities are normalized according to the author's intention, with the same *Fusarium oxysporum f.sp. cubense tropical race 4* identifier for both entities.

Example

(111) Fusarium oxysporum F. sp. Cubense, Fusarium odoratissimum ID 2502994
also known as race four tropical

Normalisation with 2 or more identifiers

Normalization also allows 2 or more identifiers to be associated with a single entity. This means the conjunction of identifiers should be used when, for example, an identifier excludes organisms to which the entity refers. As in the example (112) of the *Foc* entity annotation. NCBI identifier 61366 *Fusarium oxysporum f. sp. Cubense* includes all tropical races except 4, which is named *Fusarium odoratissimum*. The conjunction of the two is therefore necessary to standardize *Foc*.

Example

	Correct	Incorrect
(112)	<ul style="list-style-type: none"> ID 61366 (<i>Fusarium oxysporum f. sp. cubense</i>) ID 2502994 (<i>Fusarium odoratissimum</i>, homotypic synonym: <i>Fusarium oxysporum f. sp. cubense tropical race 4</i>) 	<ul style="list-style-type: none"> ID 61366 (<i>Fusarium oxysporum f. sp. cubense</i>)

3.2. Normalisation of dissemination pathways

Dissemination pathway entities are normalized by class identifiers of the OntoBiotope ontology. The version used is October 2021 at

<https://agroportal.lirmm.fr/ontologies/ONTOBIOTOPE>

Examples

(113) Water environment water OBT 000047

- (114) **Coniferous wood** ● wood OBT : 001383
- (115) **Wood Packaging Material** ● wooden pallet OBT : 003753
- (116) **nonhost species** ● living organism OBT:000010
- (117) **farm equipment** ● agricultural equipment OBT:000032
- (118) **Citrus fruits** ● citrus fruits OBT :002971 OBT:002971
- (119) **contaminated tools** ● agricultural tool OBT :000185

3.3. Normalisation of location

Locations are normalized by the identifiers of the 2022 version of GeoNames:

<https://www.geonames.org/>

Selection of the Geonames identifier is done with respect to the *Feature class* value. By default, the correct *Feature class* is the *administrative division* and the *political entity* for the country.

To choose the right level of *administrative divisions* (first to fourth order), preference is given to identifiers that represent polygons (surfaces) while *seat* or *populated place* features are discarded.

The general rules and various exceptions are described below, with illustrative examples:

General rules

- **Continent** entities → *Feature class* “continent”
- **Countries** → *Feature class* “independent political entity” or “dependent political entity” in case of disputed territory (ex. *Puerto Rico*).
- **Administrative regions** → “first / second order administrative division” (depending on the administration level of the country)
- **Towns or municipalities** → *Feature class* “second / third / fourth-order administrative division” (depending on the administration level of the country)
- If the location is not in Geonames, it is annotated by the identifier of the most specific surrounded region (examples (120) et (121)).





Examples

-
- (120) **central-southern Puglia** ● Puglia Geonames ID 3169778

(121) **western Mediterranean basin**  Mediterranean Basin Geonames ID 12217088

- Locations should not be normalized with identifiers which *Feature classes* are “*seat of a X-order administrative division*” that are the locations of the administrative seats of the corresponding location, neither with identifiers which *Feature classes* are “*populated place*” (examples (122) and (123)).

Examples


		Correct	Incorrect
(122)	city of Freiburg	 Freiburg im Breisgau Geonames ID 6555728 (fourth-order administrative division)	 Freiburg im Breisgau Geonames ID 2925177 (seat of a second-order administrative division)
(123)	Mogro	 Miengo Geonames ID 6360690 (third-order administrative division)	 Mogro Geonames ID 3116630 (populated place)

Exceptions to administrative divisions

- **Region**

Places that correspond to a particular geographical area without administrative designation can be annotated with the feature class "region" (example (124)).



Example

(124) **Central Italy**  Central Italy Geonames ID 12089032 (Feature class: region)

- **Zone**

When the entity designates a location unified by a political agreement such as the "*European Union*", it must be annotated with the feature class "*zone*" (example (125)).

Example

		Correct	Incorrect
(125)	European Union	 European union Geonames ID 6695072 (zone)	 Europe Geonames ID 6255148 (continent)

- **Physical geography**

If the entity is an element of physical geography (e.g. mountain, lake, port) that does not correspond to an administrative division, then another *Feature class* value can be selected.

Example

Correct


(126) **Rwenzori Mountains**  Massif du Ruwenzori Geonames ID 206041 (mountains)



- **Union of identifiers**

If the entity designates a region made up of several administrative divisions and it does not exist as such in Geonames, then the entity is annotated either by all the identifiers of the divisions included if they are identifiable (examples (127) and (130)), or by the most specific administrative division or zone/region encompassing it. Standardization with more than one Geonames identifier means the union of these identifiers.

Examples

Correct

(127) **Iberia**  Espagne Geonames ID 2510769
 Portugal Geonames ID: 2264397

(128) **Latin America**  South America Geonames ID 6255150
 Central America ID 7729892



- **More than one relevant identifier**

In the case where it is possible to normalize an entity with several Geonames identifiers and the polygons of these identifiers coincide (see map in Geonames), the rule indicates that we must give priority to the "administrative division" (example (129), unless the entity contains a physical geography term that characterizes the entity, such as island (example (130)).

Examples

Correct

Incorrect

(129) **Corsica**  Corse Geonames ID 3023519 (first-order administrative division)  Corse Geonames ID 3023517 (region)

(130)

Corsican island

● Corsican island
Geonames ID 3023518
(island)

● Corse Geonames ID 3023519
(first-order administrative
division)



4. Annotation of relationships

Relationships are annotated only when they are made explicit in the text. If the relationship is well-known, but not explicit, it is not annotated. The entities linked by a relationship are called the arguments of the relationship. Relationships are said to be binary when they link two arguments. They are said to be ternary for three arguments, and quaternary for four. More generally, relations with more than two arguments are called n-ary.

Distance

A given entity is linked to another entity if they are at most two sentences before and two sentences after the entity sentence. More distant entities are not linked, despite their interest, unless they are part of the main title. n-ary relations can extend over more than two sentences, from near to near.

Type of relation

There are two types of relations: (1) equivalence relations that link two entities that mean the same thing. These are called coreference relations. (2) Thematic relations link entities according to their semantic role in the interaction.

4.1. Coreference

The coreference relationship links two or more terms that refer to the same entity. It represents an equivalence between the mentions ((131) to (134)).

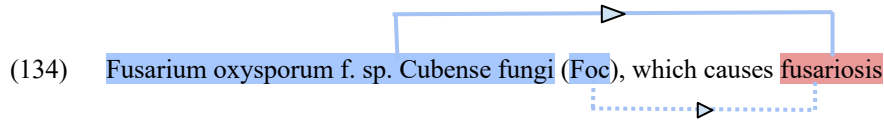
Examples

(131)	coreference between disease mentions	Panama disease, also called banana wilt, a devastating disease of bananas
(132)	coreference between pest mentions	Apple proliferation phytoplasma (APP) (Candidatus Phytoplasma mali) is a plant pest
(133)	coreference between location entities	the virus could have been spread from the Democratic Republic of Congo (DRC)

Annotation simplification by coreferences

When two entities E1 and E2 are linked by a coreference relationship, each entity inherits the thematic relationships of the other entity. It is then not necessary to annotate for E1 the relationships already involving E2.

In the example below, *Fusarium oxysporum f. Sp. Cubense fungi* and *Foc* are linked by a coreference relationship. It is not necessary to create a causal relationship between the *Foc* pathogen and the *fusariosis* disease. The relationship between *Fusarium oxysporum f. Sp. Cubense fungi* and *fusariosis* and the co-referential relationship are sufficient to establish that *Foc* and the disease *fusariosis* are also related.



When the coreferent term is not synonymous with the antecedent, the coreference relationship is not annotated, for example if the coreferent is more general. For example, "bacteria" or "nematode" are not annotated as coreference in the examples (135) and (136).

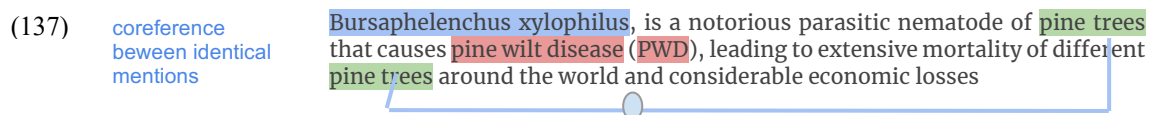
Examples



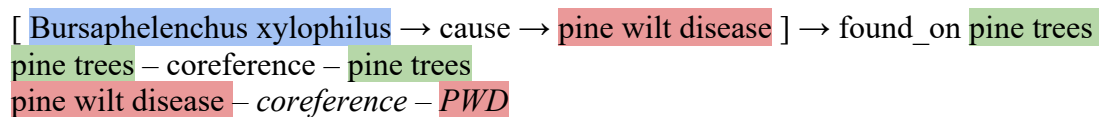
(135) Apple proliferation phytoplasma, this bacteria

(136) B. xylophilus. This nematode is widespread

Coreference can be used to link entities with a different surface form with the same meaning (e.g (134) to (137)) but also to link entities with the same surface form to have them inherit the relationships of one entity to another, provided they have an equivalent role in the argument (example (137)):



In the example (140), the relationships are:



4.2. Thematic binary relationship

The thematic binary relationship links two entities. It is oriented, and it has a label. The entity with the starting argument is called the source, and the entity with the ending argument is called the target. Its label characterizes the type of relationship - for example, *cause* between *pest* and *disease*. There are 8 different types.

The sections below are organized by binary relationship type. For each, the type of possible source and target entities are indicated. The table below summarizes the argument types and labels.

Table 1. Valid relationships with respect to the argument types.

->	Pest	Plant	Vector	Disease	Dissemination _pathway	Date	Location
Pest	-	Found_on	Vected_by	Causes	Found_on Dispersed_by	Detected_ by_	Located_in
Plant	-	-	-	-	-	Detected_ by_	Located_in
Vector		Found_on	-	-	Found_on	Detected_ by_	Located_in
Disease	-	Expressed_ by_	-	-	Dispersed_by	Detected_ by_	Located_in
Dissemination_ pathway	-	-	-	-	-	Detected_ by_	Located_in
Date	-	-	-	-	-	-	-
Location	-	-	-	-	-	-	-

4.2.A. Causes

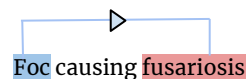
Source entity **Pest**

Target entity **Disease**

Definition Relation between a pest and the disease that it causes.

Example

(138)



4.2.B. Found_on

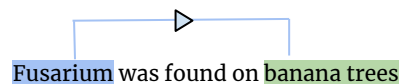
Source entity **Pest** or **Vector**

Target entity **Plant** or **Dissemination_pathway**

Definition Relation between a pest or a vector and where it can be found, plant or dissemination pathway.

Example

(139)



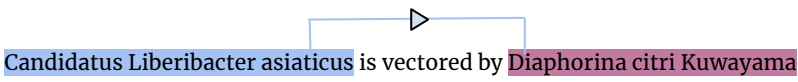
4.2.C. Vected_by

Source entity **Pest**

Target entity **Vector**

Definition Relation between a pest and a vector.

Example

(140)  **Candidatus Liberibacter asiaticus** is vectored by **Diaphorina citri Kuwayama**

The diagram shows a blue line connecting the two entities. The line starts from the top of 'Candidatus Liberibacter asiaticus', goes up, then right, then down, then right, then down, ending with a triangle arrowhead pointing to the top of 'Diaphorina citri Kuwayama'.

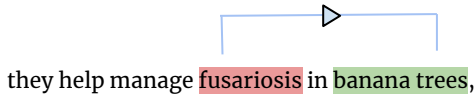
4.2.D. Expressed_by

Source entity **Disease**

Target entity **Plant**

Definition Relation between a plant and the disease that it expresses.

Example

(141)  they help manage **fusariosis** in **banana trees**,

The diagram shows a blue line connecting the two entities. The line starts from the top of 'fusariosis', goes up, then right, then down, then right, then down, ending with a triangle arrowhead pointing to the top of 'banana trees'.

4.2.E. Dispersed_by


Source entity **Pest** or **Disease**

Target entity **Dissemination_pathway**

Definition The relationship between the pest or disease, and its dissemination pathway. In fact, diseases are not dispersed in themselves, but are often assimilated to the disease agent in texts.

Example

(142) common pathway for **Oriental fruit fly** to enter the State is by “hitchhiking” in **fruits**



4.2.F. Located_in


Source entity **Pest** or **Vector** or **Plant** or **Disease** or **Dissemination_pathway**

Target entity **Location**

Definition Relation between the observed entity and the observation location.
A pest that causes an epidemic is located in the epidemic location.

Example

(143) **Fusariosis** is common in **Brazil**



4.2.G. Detected_by


Source entity **Pest** or **Vector** or **Plant** or **Disease** or **Dissemination_pathway**

Target entity **Date**

Definition Relation between the observed entity and the observation date.

Example

(144) **HLB** was reported in **South China** in **1943**.



4.3. N-ary relationships (events)

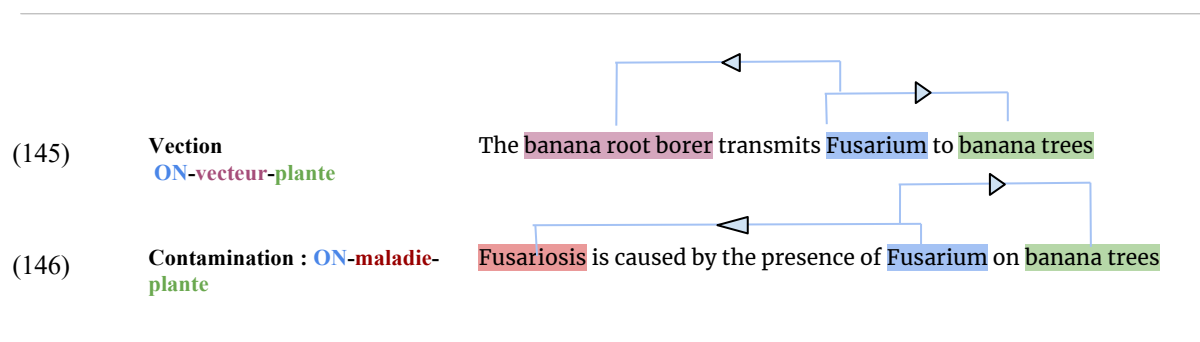
Definition

N-ary relations are relations that link more than two arguments. They are also known as "events". A first binary relation links the two main entities. The other relations link the first relation with the additional entities. The meaning of the relationship is the same as for binary relationships

The relationships have a fixed order of priority:

1. *Causes* [Pest -> Disease]
2. *Found_on* [Pest -> Plant]
3. *Vected_by* ou *Dispersed_by* [Pest or Disease -> Vector or Dissemination_pathway]
4. *Located_in* [? -> Location]
5. *Detected_by* [? -> Date]

Examples



4.4. Modality

Modality feature is associated to some relationships. It can be neural, by default, negation or hypothesis. This modality can be used when a resistance relationship is described between the host plant and the pest.

4.4.A. Negation

The Modality feature takes the value « negation » when the relationship is explicitly negated in the document.

Example



- (148) *negation* Cavendish variety with complete resistance to Panama disease will be available in 2024
- (149) *negation* Formentera is the only island, at the moment, free of Xylella fastidiosa
- (150) *negation* in the framework of the 2022 survey program for the harmful organism xylella fastidiosa in its host plants (after relevant consultation with the Benaki Plant Pathological Institute) in order to confirm the absence of the harmful organism (detection survey).
-

4.4.B. Hypothesis

The Modality feature takes the value « hypothesis » when the relationship is hypothetical in the document.

Example

- (151) *hypothesis* Crotalaria [...] identified as potential hosts of Foc
- (152) *hypothesis* Gathering direct evidence about whether the bacteria infect pepper tissues may prove difficult, as Hansen's team only had a single sample, and L. capsica cannot be grown in a laboratory.
-