Guidelines for the annotation of the Taec corpus - phenotype and trait information in wheat

D2KAB project

Version 1. 5 May 2022
Author : Claire Nédellec, Louise Deléger, Clara Sauvion

Based on
Claire Nédellec, Robert Bossy, Marion Ranoux, Pierre Sourdille, Dialekti Valsamou, *Guidelines for the annotation of information for marker-assisted selection in wheat FSOV Sam Blé project*. 2013-04-12. https://hal.science/hal-03620421v1

This document presents the guidelines for the manual annotation of trait and phenotype mentions of the *Triticum aestivum Trait Corpus* (*Taec)* https://doi.org/10.57745/GCYG3Q

Guidelines for the annotation of phenotype and trait information in wheat

0 Note

This document is an extended version of the annotation guidelines of the FSOV SamBlé project. It includes the addition of the phenotype entity and the normalization of the phenotype and trait entities. The rest of the entity and relations of SamBlé guidelines are not considered here.

1 Introduction

This document specifies the guidelines for the annotation of the Wheat Trait and Phenotype D2KAB corpus. The task consists of the extraction of plant species, traits, and phenotypes of bread wheat varieties in a set of scientific texts (Pubmed abstracts). Species are annotated by the NCBI taxonomy entries, trait, and phenotype entities are normalized by classes from the Wheat Trait and Phenotype Ontology.

1.1 Copyright and License

1.2 Conventions

In the examples, `Trait` annotations are highlighted in light green, `Phenotype` annotations in dark green. Species annotation are highlighted in turquoise .

2. Mentions

2.1 Trait

2.1.1 Trait definition

The traits refer to the observable characters or properties but do not include the phenotype, which is the observable value of the trait.

Examples

In tests for resistance to P. triticina race 5, plants

wheat cultivars to provide protection from WSM

2.1.2 Boundaries

The trait mention may include the name of species that expresses the trait

*resistance to **Wheat** streak mosaic virus (WSMV)*.

or of the pathovar

*In tests for resistance to P. triticina **race 5**, plants*

The trait includes the properties of the plant that condition the trait conditions for the trait

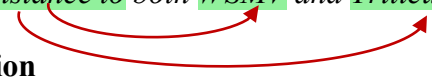
High—temperature adult-plant (HTAP) resistance to stripe rust of wheat.

genes that confer seedling resistance in Chinese wheat cultivars

Discontinuous annotation

Distinct entities must be annotated separately.

resistance to both WSMV and Triticum mosaic virus



Apposition

A trait mention that includes appositions must be annotated in a single fragment (not discontinuous) if it is part of the mention.

current knowledge about genes for resistance to Septoria tritici blotch (STB) of wheat

If the apposition is a distinct mention, it is annotated separately.

Plant height (PHT) is a crucial trait related to plant architecture

2.1.3 Trait domain

Mentions of traits are expressions and phrases that denote a characteristic of the plant. This includes:

- morphology of the plant or part of the plant (color, size, etc.);
- response to biotic and abiotic stress (tolerance, damage, etc);
- development (flowering time, growth habit, etc.))
- quality (grain hardness, starch content, etc.)

This excludes:

- Diseases, symptoms, pests
- Diagnostic or observation methods
- Environmental conditions (temperature, humidity, etc.)

When the trait mentioned characterizes an organism that is not a plant, but is applicable to, it is annotated.

improve the viability of E. coli under heat and cold stress

increase the Saccharomyces cerevisiae tolerance under salt and osmotic stress.

2.1.4 Over general trait mention

When a trait mention is too generic or too imprecise, it must not be annotated. The following list is a vocabulary of terms that are too generic:

- “trait”
- “character”
- “resistance” without any further specification

2.2 Phenotype

2.2.1 Phenotype definition

The phenotype is the value of the trait.

Examples

cultivars that originally were resistant to leaf rust

Eight dwarf and semi-dwarf varieties, covering a range of genetic sources

2.2.2 Boundaries

The phenotype mention may include the **name of pathogen species** that expresses the trait if it is part of the phenotype name

*resistant to **Wheat** streak mosaic virus (WSMV).*

or of the **pathovar**

*accessions that were resistant to the **Ug99 race group***

If the pathogen name is overgeneral or too vague to identify the species, then it is not annotated

*accessions that were resistant to the **Type I and II***

The phenotype mention includes the **properties of the plant** that condition the trait conditions for the trait

*wheat cultivar Kariega expresses complete **adult plant** resistance against stripe rust*
*for **field** resistance to the Ug99 stem rust pathogen*
*The Yr18 gene is known to confer slow rusting resistance **in adult plants***
*resistant at the **adult plant stage***

Counter-example

*The parental accessions were susceptible to all the prevalent pathotypes **at the seedling stage**,*

The word Phenotype itself should not be included

*in relation to the lodging-resistant **phenotype** in wheat*

Discontinuous annotation

Distinct entities must be annotated separately.

resistant to both WSMV and Triticum mosaic virus

It happens that the phenotype term does not include the name of the trait. However, it is annotated.

and highly-susceptible cultivar Wheaton

Apposition

A phenotype mention that includes appositions must be annotated in a single fragment (not discontinuous).

evaluation of fusarium head blight resistant (FHB) wheat germplasm

2.2.3 Phenotype domain

Mentions of phenotypes are expressions and phrases that denote a value for the characteristic of the plant. This includes:

- the morphology value of the plant or part of the plant (white as a color, small as size, etc.);
- the response value to biotic and abiotic stress (resistant, highly susceptible, etc.);
- development value (winter habit, etc.)
- quality (shriveled grain, etc.)

This excludes:

- Diseases, symptoms, pests
- Results of diagnostic or observation methods
- Values of environmental conditions (high temperature, low humidity, etc.)

The phenotype must not be confused with the trait, or the environmental factor.

Photoperiod has an important effect on plant growth

Photoperiod is not a phenotype of the plant, but a factor of the environment.

2.2.4 Over general phenotype mention

A phenotype mention must not be annotated when it is too generic or imprecise. Single adjectives or adverbs denoting a value should not be annotated. High, low, and extremely are examples. The following list is a vocabulary of terms that are too generic:

- "phenotype"
- "value"

2.3 Taxon

2.3.1 Taxon definition

Mentions of a taxon are expressions and phrases that denote a name of a plant. This includes:

- scientific names;
- vernacular names;
- infra species when defined in the reference NCBI taxonomy

(https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?).

This excludes Varieties or Cultivars that are created for experimental purposes.

The bread taxon (*Triticum aestivum*) annotates wheat and common wheat mentions by default.

improve the efficiency of designed breeding in wheat

Durum mention is annotated by the durum wheat taxon *Tritucum turgidum* subsp. by default.

Tunisian durum landraces

the short arm of chromosome 2D of common wheat (Triticum aestivum L.).

2.3.2 Over general trait mention

A taxon mention must not be annotated when it is too generic or imprecise. The following list is a vocabulary of terms that are too generic:

- "plant"

2.3.3 Boundaries

The name of the taxon should include the binomial scientific name

Triticum aestivum

The name of the taxon should include the binomial scientific name, the authority, and the year when specified

Triticum aestivum L., 1753

The name of the taxon includes the variety, the subspecies, and the crossing when defined in the NCBI taxonomy

Triticum aestivum var. lutescens
Triticum aestivum subsp. hadropyrum
Agropyron x *Elymus*

The vernacular name of a taxon includes the adjective that is necessary to specify the taxon but excludes unnecessary modifiers.

resistant bread wheat varieties
synthetic hexaploid wheat

3. Normalization

3.1 Normalization of traits and phenotypes

Trait and phenotype mentions are associated to one relevant class of the Wheat Trait and Phenotype Ontology, version 2.2 through the attribute WTO.

In simple cases, the label or the synonym of the WTO class is close to the trait or phenotype mention.

*resistance to **Wheat** streak mosaic virus (WSMV)* is associated to
WTO:0000568 'resistance to wheat streak mosaic virus'

In more complex cases, the terms may strongly differ.

When no phenotype class can be associated to a phenotype mention, then the mention is associated to the trait that corresponds to the phenotype.

Reduced starch content accounts for most of the reduction in grain dry matter at high temperature.
Reduced starch content -> WTO:0000131 "grain starch content"
reduction in grain dry matter -> WTO:0000131 "dry matter yield"

3.2 Normalization of taxa

Taxon mentions are associated with one relevant class of the NCBI taxonomy (https://www.ncbi.nlm.nih.gov/Taxonomy/Browser/wwwtax.cgi?).
In simple cases, the label or the synonym of the NCBI taxon is close to the taxon mention.

Triticum aestivum is associated to the Taxonomy ID: 4565 ' Triticum aestivum
In other cases, the mention is a synonym. *e.g.,* bread wheat, an abbreviation, or an acronym.