Taec : Un corpus annoté pour l'extraction de traits et phénotypes dans la littérature sur la sélection du blé

Claire Nédellec, Clara Sauvion, Robert Bossy, Louise Deleger

Taec : Un corpus annoté pour l'extraction de traits et phénotypes dans la littérature sur la sélection du blé

Claire Nédellec, Clara Sauvion, Robert Bossy, Louise Deléger

[1] Université Paris-Saclay, INRAE, MaIAGE, Jouy-en-Josas, France

Corresponding author: Claire Nédellec
E-mail: claire.nedellec@inrae.fr

Résumé

Les variétés de blé présentent une très grande diversité de traits et de phénotypes. Il est essentiel de les relier à la variabilité génétique pour rendre les programmes de sélection du blé plus courts et plus efficaces. Parmi les nouvelles caractéristiques souhaitables des variétés de blé figurent la résistance aux maladies afin de réduire l'utilisation de pesticides, l'adaptation au changement climatique, la résistance aux stress liés à la chaleur et à la sécheresse, ou la faible teneur en gluten des grains. De nombreuses informations sur la relation génotype-phénotype sont décrites dans la littérature scientifique sur la sélection assistée par marqueurs. La diversité des termes utilisés pour désigner les caractères et les phénotypes dans les articles constitue un obstacle à la recherche d'informations et à leur recoupement avec les observations phénotypiques obtenues sur le terrain et *in vitro*.
Lorsqu'elles sont correctement entraînées à partir d'exemples annotés, les méthodes récentes de traitement automatique de la langue (TAL) atteignent des performances élevées en matière de reconnaissance d'entités nommées et de liage (normalisation). Alors que des corpus contiennent des annotations de phénotypes humains et animaux, il n'existe pas de corpus sur la littérature relative aux phénotypes végétaux permettant d'entraîner et d'évaluer les méthodes de reconnaissance et de liage d'entités nommées. Le corpus *TaeC* est un nouveau corpus de référence pour les traits et les phénotypes du blé. Il se compose de 528 références de la base bibliographique PubMed entièrement annotées pour les traits, les phénotypes et les espèces à l'aide de l'ontologie des traits et des phénotypes du blé WTO et de la taxonomie des espèces du NCBI. Ce corpus est actuellement la ressource la plus complète en matière d'informations sur les phénotypes végétaux.

Introduction

L'amélioration de beaucoup d'espèces végétales d'intérêt agronomique dans un avenir proche est devenue un enjeu international en raison de la demande croissante pour nourrir une population mondiale en pleine croissance et pour atténuer la réduction des ressources, en particulier en eau et en énergie. Ceci est particulièrement vrai pour le blé, la culture la plus répandue dans le monde après le riz. Le changement climatique et la réduction des intrants (eau, engrais et pesticides) et des surfaces cultivées sont de nouvelles contraintes environnementales qui entraînent un besoin accru de variétés de blé présentant des caractéristiques liées à la tolérance à la limitation de l'eau et au stress thermique, à la résistance aux effets de la vitesse du vent (par exemple, résistance mécanique de la tige, résistance à la verse), à la résistance aux maladies ou à l'efficacité de l'utilisation des nutriments (Paux et al., 2022). Le blé est utilisé dans une large gamme de nouveaux produits alimentaires, dont la demande augmente en raison de l'évolution des régimes alimentaires et de la recherche de bienfaits pour la santé. La composition du grain de blé est un objectif majeur de changement : par exemple, préparations à faible teneur en gluten, protéines texturées pour les produits végétariens, composition de l'amidon intervenant dans la préparation de produits de boulangerie, de produits carnés et de confiseries, taux plus faible de produits chimiques synthétiques dans la composition des panures, des enrobages et des additifs de saumure.
Les caractéristiques du blé qui font l'objet de recherches sont donc très diverses, allant de la réponse aux conditions environnementales (biotiques et abiotiques), à la qualité pour l'alimentation, à la croissance (par exemple, le rendement, la vigueur, l'utilisation des nutriments), à la morphologie et à la reproduction.
Les progrès récents des outils génomiques ont contribué à améliorer le lien entre les marqueurs moléculaires et les gènes d'intérêt agronomique. Ces informations sont intégrées dans des programmes de sélection de plus en plus courts dans le but de passer d'une sélection génétique à une sélection génomique. Un grand nombre de variétés et de marqueurs moléculaires ont été développés ces dernières années pour le blé panifiable (Tadesse et al., 2019). Les données des

expériences de sélection sont réparties dans des milliers d'ensembles de données et de publications hétérogènes (Ćwiek-Kupczyńska et al., 2016). L'objectif du projet *D2KAB* (*Data to Knowledge in Agriculture and Biodiversity* ; https://d2kab.mystrikingly.com/), dont ce travail fait partie, est de développer de nouveaux outils basés sur le web sémantique pour la description sémantique des données agronomiques, en les rendant exploitables et ouvertes, selon les principes FAIR (findable, accessible, interoperable, and reusable) (Wilkinson et al., 2016). Les sélectionneurs utilisent la Crop Ontology pour annoter les données d'observations phénotypiques dans le but de relier les données phénotypiques et génétiques pour la sélection intégrée (Cooper et al., 2018). Cependant, l'annotation automatique des données textuelles reste un défi en raison du grand nombre de traits et de la grande diversité des termes pour les désigner. En outre, les traits dans les documents qualifient principalement les propriétés générales des variétés, tandis que les données d'observation qualifient des états donnés de la plante dans un champ spatial et temporel limité qui doivent être agrégés et confirmés expérimentalement pour dériver les propriétés générales de la variété de blé observée. L'ontologie des traits et des phénotypes du blé (WTO ; http://agroportal.lirmm.fr/ontologies/WHEATPHENOTYPE) a été développée pour répondre aux besoins de l'annotation de données textuelles par ontologie (Nédellec et al., 2020), c'est-à-dire le liage des entités nommées (LEN). *WTO* contient 502 classes de traits et de phénotypes et en couvre toutes les dimensions.
Comme le montre le tableau 1, les labels des classes de l'ontologie et des termes de l'article diffèrent à bien des égards, ce qui empêche l'application d'une méthode de correspondance directe des chaînes de caractères pour le LEN. La même remarque s'applique à CO_321, la Crop Ontology dédiée au blé tendre (*Triticum aestivum*). Des méthodes d'extraction d'informations plus puissantes sont nécessaires pour automatiser la reconnaissance des entités de traits et de phénotypes et les relier aux classes pertinentes de l'ontologie WTO.

Table 1. Exemples à lier d'ontologie et de mentions de texte.

Text entity	**Class label**
heading time	Ear emergence time
number of fertile florets at anthesis	percentage of florets without grain
highly resistant to leaf rust caused by *Puccinia triticina*	resistance to Leaf Rust
resistance to single isolates of *M. graminicola*	resistance to Septoria Leaf Blotch
HTAP resistance	high-temperature resistance
phenolic content	grain polyphenol content
number of tillers	shoot number per plant
TKW	thousand kernel weight
low molecular weight glutenin subunit	glutenin content

Des corpus annotés sont nécessaires pour évaluer les méthodes d'extraction d'informations. Bien que le nombre d'exemples d'apprentissage requis pour entraîner les méthodes basées sur l'apprentissage automatique diminue avec l'avènement des méthodes "few-shot" et "zero-shot" (Sevgili et al., 2022), la qualité de la prédiction reste corrélée à la disponibilité d'exemples d'apprentissage de l'information cible.
Les corpus pour la reconnaissance d'entités nommées (REN) et le liage d'entités nommées (LEN) en sciences de la vie se concentrent principalement sur la santé humaine. Peu d'entre eux contiennent des annotations sur les propriétés des plantes, et aucun ne contient d'annotation de traits par ontologie. Nous avons conçu le corpus des traits sur le blé tendre, *Triticum aestivum Trait Corpus* (*TaeC*) pour combler cette lacune. Nous avons choisi *PubMed* (https://www.ncbi.nlm.nih.gov/pmc/) comme source bibliographique de documents scientifiques parce qu'elle est entièrement ouverte, que les titres et les résumés des références sont des textes courts et ciblés, et qu'ils contiennent une grande quantité de mentions de traits et de phénotypes. Ces entités sont liées dans *Taec* aux classes de l'ontologie WTO. Outre les traits et phénotypes, les types d'entités incluent les espèces pour lesquelles nous avons utilisé la taxonomie du NCBI comme référence sémantique.

Travaux connexes

La plupart des travaux d'extraction d'informations se sont concentrés sur les phénotypes humains et animaux et leur lien avec les particularités et les anomalies génétiques. Le corpus populaire *Phenotype-Gene Relations* (PGR) (Sousa et al., 2019) a été annoté par l'ontologie de phénotype humain, *Human Phenotype Ontology* (HPO), qui est un vocabulaire standard d'anomalies phénotypiques rencontrées dans les maladies humaines (Köhler et al., 2017) (8), à l'aide de l'outil de REN basé sur l'apprentissage automatique IHP (*Identifying Human Phenotypes*) (Lobo et al., 2017). HPO n'inclut pas les traits réguliers tels que la couleur des yeux, mais seulement les traits anormaux tels que l'albinisme oculaire. Le corpus *Bacteria Biotope* (Bossy et al., 2019) a été annoté par l'ontologie *OntoBiotope* (Nédellec et al, 2017), un vocabulaire standard sur les biotopes et les phénotypes des microbes. Bien qu'*Ontobiotope* décrive des phénotypes normaux, les microbes sont trop différents des plantes pour que le corpus et l'ontologie soient réutilisables, même si les classes de haut niveau sont en partie les mêmes, par exemple la réponse au stress biotique et abiotique, ou la morphologie.
Jusqu'à présent, la biologie végétale a été relativement peu représentée pour la communauté de traitement de la langue en domaine biomédical *BioNLP*. La plante *Arabidopsis thaliana* a récemment fait l'objet de quelques initiatives dans le domaine de l'extraction d'information, telles que le système de curation de la littérature *KnownLeaf* (Van Landeghem et al., 2013 ; Szakonyi et al., 2015) et le corpus de référence *SeeDev* (Chaix et al., 2016). Le corpus *SeeDev* se concentre sur le développement des graines décrit au niveau moléculaire. Le corpus *Knownleaf* se concentre sur les mécanismes de régulation de la croissance et du développement des feuilles, et sur les gènes clés en relation avec les phénotypes mutants pertinents. Les traits végétaux du corpus *Knownleaf* ont été décrits à l'aide de termes de l'ontologie des phénotypes, des attributs et des caractères, *Plant Attribute and Trait Ontology* (*PATO*) (Smith et al., 2007), qui sont combinés aux parties et tissus des plantes de l'ontologie des traits de plante (*TO*) (Liang et al., 2007) et de l'ontologie des tissus *Brenda* (Gremse et al., 2010), selon le modèle entité-attribut-valeur (EAV). Par exemple, dans le texte *The reduced leaf area in the hub1-1 mutant*, le phénotype *reduced leaf area* (surface réduite de la feuilles) est formalisé en trois entités distinctes, *area* (surface) comme propriété, *reduced* (réduite) comme valeur, et *leaf* (feuille) comme partie de la plante. Notre objectif diffère de celui de *Knownleaf* dans la mesure où nous ne considérons pas les phénotypes du blé comme normaux ou anormaux par rapport à une référence génétique normale, mais nous visons à rendre compte de toute la diversité des valeurs des caractères dans les variétés de blé. Pour *TaeC*, nous avons également préféré les annotations textuelles qui couvrent à la fois la partie de la plante, le caractère et potentiellement sa valeur, plutôt que la distinction formelle des entités comme le fait le projet *Knownleaf,* par exemple, *grain color* (couleur du grain). L'objectif du projet *D2KAB* est en effet de faciliter la recherche et la lecture d'informations par des utilisateurs humains. L'annotation des données textuelles et expérimentales est ensuite réalisée en utilisant WTO pour le texte, et CO_321 pour les expériences. Dans les deux ontologies, les classes de traits regroupent à la fois le trait et la partie de la plante qu'il caractérise ; par exemple, *grain color* est le label des classes de traits WTO:0000141 et CO_321:0000037. La stratégie d'annotation et la structure des labels des classes sont donc cohérentes.

Matériel et Méthodes

Dans cette section, nous décrivons en détail comment nous avons construit le corpus de référence *TaeC*. Nous présentons notre méthode de sélection des documents, le schéma d'annotation des données, les consignes d'annotation et le processus d'annotation.

Sélection des documents

Les documents sont sélectionnés par la requête exécutée sur PubMed (Table 3).

Table 3. Requête de sélection de documents sur PubMed.

(("biomarkers"[MeSH Terms] OR "biomarkers"[All Fields] OR "marker"[All Fields] OR "markers"[All Fields]) *AND ("genes"[MeSH Terms] OR "genes"[All Fields] OR "gene"[All Fields])* *AND ("triticum"[MeSH Terms] OR "triticum"[All Fields] OR "wheat"[All Fields] OR "wheat s"[All Fields] OR "wheats"[All Fields]))* *AND ((fha[Filter]) AND (english[Filter]))*

La première partie de la requête (3 premières lignes), avec les mots-clés marqueur et gène, sélectionne les documents sur la sélection génétique du blé et exclut les documents sur d'autres utilisations du blé telles que la transformation ou la composition de produits alimentaires. Les lignes 4 et 5 de la requête sélectionnent des documents sur les espèces de blé. La dernière partie de la requête sélectionne les documents en anglais.
La requête effectuée en 2021 a permis de récupérer 5 596 documents qui ont été publiés pour la plupart après 2011. Afin de sélectionner un sous-ensemble représentatif à annoter, nous avons appliqué le workflow *AlvisNLP* dédié aux REN et LEN du blé tendre pour annoter les traits et phénotypes dans les titres et les résumés et les lier aux classes de *WTO* (Nédellec et al., 2014). Nous avons ensuite généré des sous-ensembles de 6 documents où la sélection est biaisée de façon à ce que chaque ensemble contienne au moins un document avec des entités annotées par l'un des six sous-arbres principaux de *WTO* : développement, croissance, morphologie, qualité, reproduction et réponse aux conditions environnementales. Un document peut être annoté par plus d'un sous-arbre.

Schéma d'annotation

Les types d'entités de *TaeC* comprennent les types trait, phénotype et espèce. Les traits sont des caractéristiques observables telles que la hauteur de la plante ou la résistance à une maladie donnée. Les phénotypes sont les valeurs des traits, par exemple 1,2 m comme valeur de la hauteur (*plant height*), ou *hautement résistant à la pyriculariose du blé* (*highly resistant to wheat blast*) comme valeur du trait de *résistance à la pyriculariose du blé (resistance to wheat blast)*. Nous avons utilisé WTO comme ressource de référence pour leur annotation.
Les espèces sont des entités pertinentes ici puisque les variétés hybrides obtenues par hybridation avec des plantes sauvages (par exemple, *Aegilops tauschii*, une espèce de graminée annuelle) et cultivées (par exemple, *Hordeum vulgare*, l'orge) sont mentionnées en plus du blé tendre (*Triticum aestivum*), du blé dur (*Triticum durum*), et de leurs sous-espèces. Nous avons utilisé la taxonomie du *NCBI* (https://www.ncbi.nlm.nih.gov/taxonomy) comme ressource de référence pour l'annotation standard des espèces. La taxonomie du *NCBI* est une ressource mondialement reconnue qui est également pertinente pour l'association des gènes aux phénotypes grâce à ses bases de données de séquences génétiques liées.
Les types d'entités de *TaeC* n'incluent pas les variétés ou les cultivars malgré leur pertinence. Cette question est laissée à des travaux futurs en raison de sa grande complexité. Les variétés peuvent certes être mentionnées par leur nom commercial (référencées par exemple dans *European Plant variety database* ; https://ec.europa.eu/food/plant-variety-portal/), ce qui les rend triviales à reconnaître mais sans grande valeur pour *TaeC*. Les mentions les plus fréquentes et les plus utiles pour un corpus de référence sont celles des cultivars ou des accessions obtenues par modification génétique ou par croisement. Malheureusement, il ne s'agit souvent pas d'entités nommées, mais d'expressions plus complexes, y compris verbales. Le tableau 2 présente quatre exemples illustratifs de ces expressions. Leur annotation manuelle requiert un niveau élevé d'expertise et de ressources à un coût élevé.

Table 2. Exemples de noms et descriptions de cultivars.

[..] F2 and F2:3 lines derived from the cross of the multi-spikelet female 10-A and the uni-spikelet male BE89 [PMID: 32549690]
[..] we produced a F(6:7) recombinant inbred line (RIL) population by crossing Wangshuibai with the scab-susceptible cultivar Nanda2419 [PMID: 15290053]

[..] wheat cultivars with Pina-null/Pinb-null allele [PMID: 24011219]

The wheat accession H9020-1-6-8-3 is a translocation line previously developed from interspecific hybridization between wheat genotype 7182 and *Psathyrostachys huashanica*. [PMID: 30727301]

Consignes d'annotation

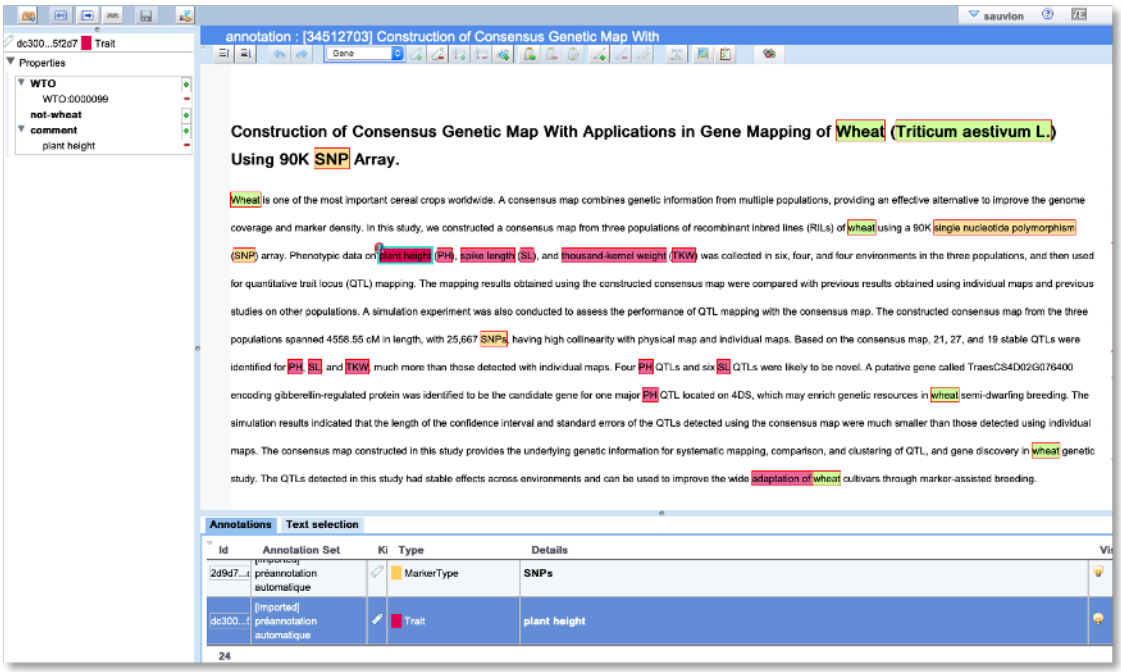
Nous avons préparé les consignes d'annotation en adaptant celles de (Nédellec et al. 2014) : nous avons supprimé les sections concernant les relations et les entités non pertinentes, et nous avons ajouté de nouvelles sections pour le liage entre les entités des traits, des phénotypes et des espèces. Il en résulte un document de huit pages qui définit les entités, donne des exemples et des contre-exemples d'annotations, et détaille les exceptions et les cas limites.

Processus d'annotation et outils

Les entités du corpus ont été pré-annotées par le workflow *AlvisNLP* dédié au blé tendre et fournies aux experts pour plus d'efficacité. Deux annotateurs ont successivement annoté 540 documents ; le premier annotateur de niveau universitaire a effectué une annotation exhaustive, puis un second annotateur expert en biologie a revu toutes les annotations. Les questions ont été traitées en collaboration avec deux experts en annotation textuelle et deux experts en caractères agronomiques du blé et en résistance aux maladies, qui étaient les caractères les plus complexes à annoter.

Pour l'annotation manuelle des textes, nous avons utilisé l'éditeur *AlvisAE* (Papazian et al., 2012) qui gère l'annotation des entités par des ontologies. Il est implémenté comme une application collaborative Web, ce qui facilite la participation et la collaboration des experts du domaine. La figure 1 montre une capture d'écran de la fenêtre principale de l'instance d'*AlvisAE* sur le blé.

Figure 1. L'éditeur d'annotation *AlvisAE* utilisé pour annoter manuellement *TaeC*.



Un contrôle de cohérence automatique a été appliqué régulièrement pour garantir une annotation de qualité. Les règles à vérifier étaient les suivantes : (1) toutes les mentions d'entités sont indexées par des classes, (2) les mentions d'entités identiques sont marquées par les mêmes classes, et (3) il n'y a pas d'entité ayant à la fois des occurrences annotées et des occurrences non annotées. Les règles peuvent ne pas être pertinentes dans certains cas, c'est-à-dire qu'une

expression donnée peut avoir des significations différentes en fonction du contexte. Le contrôle de cohérence n'a donc pas été utilisé comme un outil de révision automatique, mais était destiné à avertir les annotateurs d'erreurs potentielles.

Résultats

Dans cette section, nous présentons les détails de *TaeC*. Le tableau 4 présente les fréquences d'annotation des entités et des classes par type. La densité élevée des annotations de traits et de phénotypes par document (soit 13) confirme le choix pertinent des références PubMed. La diversité des termes est illustrée par le nombre élevé d'expressions uniques par rapport au nombre d'occurrences. Chaque mention de phénotype est répétée 1,6 fois en moyenne (1598/977) et chaque mention de trait est répétée 3,2 fois en moyenne (5453/1696). Les entités traits et phénotypes sont étiquetées par 233 classes différentes, soient 11,4 mentions uniques et 30,3 occurrences par classe en moyenne, ce qui est élevé. Comme on peut s'y attendre, la distribution varie considérablement d'une classe à l'autre. Par exemple, la classe WTO:0000072 *culm length* a 4 formes dans le corpus, *culm length*, *stem length*, et leurs acronymes CL et SL, tandis que WTO:0000146 *grain protein content* présente 45 formes différentes.
La diversité des noms de taxons est plus faible, avec 2,6 formes par classe en moyenne. Le nombre d'occurrences (6,8 par document) est étonnamment élevé, mais les occurrences sont fortement répétées (13,3 fois en moyenne), notamment le mot *wheat* avec 1595 occurrences.

Table 4. Chiffres pour *TaeC*.

# Documents	540		
# Tokens	142 726	264 par document en moy.	
# Occurrences de phénotype	1 598	2,96 par document en moy.	977 uniques
# Occurrences de trait	5 453	10.1 par document en moy.	1 696 uniques
# Classes de traits ou phénotypes	233	11 formes par class en moy.	
# Occurrences d'espèces de	3 584		
plantes	3 697	6,8 par document en moy.	278 uniques
# Taxons	106	2,6 formes par class en moy.	
# Classes de taxon	**10 635**		
# Total occurrences			

Traits et phénotypes

Les traits et phénotypes sont dénotés par des expressions adjectivales ou nominales. Beaucoup d'entre elles sont des acronymes (e.g., *GPC* pour *Grain Protein Content*); les groupes prépositionnels sont fréquents (e.g., *ratio of the quantity of glutenin to those of gliadin*, *number of grains per spike*); certaines formes incluent des expressions parenthésées (e.g., *number of flowering branches (spikelets) per node*) où le terme entre parenthèses peut être un acronyme (e.g., *Efficient phosphate (Pi) uptake*, *responses to low red light/far-red light (R/FR) ratios*). Les entités discontinues sont rares. Par exemple, dans le texte *resistant to both WSMV and Triticum mosaic virus*, deux annotations discontinues sont faites, *resistant to WSMV*, et *resistant to* Triticum *mosaic virus*.
Les expressions nominales sont les plus fréquentes, mais les phénotypes sont souvent exprimés par des adjectifs (par exemple, *sprouting-resistant*, *salt-tolerant*). L'étiquette des classes diffère des entités à bien des égards, ce qui nécessite une expertise approfondie du domaine. Tout d'abord, les parties de la plante dans les noms de traits et de phénotypes peuvent être désignées par des noms différents (par exemple, *grain/kernel/seed* ; *spike/ear* ; *culm/stem*). L'ontologie WTO définit les noms alternatifs pour ces caractères, mais pas tous. La résistance aux maladies (par exemple, *resistance to wheat blast*) peut être exprimée comme la résistance à l'agent pathogène (par exemple, *resistance to Magnaporthe grisea*). Les champignons sont la principale cause de maladies du blé. Outre leurs noms officiels, d'autres noms sont utilisés pour désigner les champignons en fonction des stades de reproduction distincts, téléomorphes, anamorphes et holomorphes. Ces différents noms sont utilisés dans les expressions

phénotypiques des textes. L'ontologie *WTO* en recense un grand nombre. Par exemple, la maladie *eyes spot* est causée par le champignon *Helgardia herpotrichoides* (syn *Pseudocercosporella herpotrichoides, Ramulispora herpotrichoides, Tapesia yallundae, Cercosporella herpotrichoides*). La classe de caractères de résistance correspondante, *resistance to eyes spot*, a donc cinq noms alternatifs.
Le nombre de noms alternatifs augmente avec le nombre d'agents pathogènes pour une maladie donnée et les connaissances sur les agents causaux évoluent avec le temps. Les experts peuvent ne pas être d'accord sur les agents causaux des maladies et sur la distinction entre les espèces. Par exemple, les synonymes de WTO:0000510 *resistance to wheat blast* comprennent *resistance to Magnaporthe grisea*, *resistance to Magnaporthe oryzae*, et *resistance to Pyricularia grisea* qui sont considérées comme des espèces différentes par certains experts, notamment ceux du *NCBI*. La contribution à l'annotation de *Taec* par des experts en maladies fongiques du blé a été significative dans le traitement de cette question.
La distinction entre le caractère et la méthode de mesure de sa valeur est une autre source d'ambiguïté. Par exemple, les caractères relatifs au poids des grains (*grain weight*) sont généralement qualifiés par la méthode, par exemple le poids de mille grains (*thousand grain weight*), le poids spécifique (*test weight*). Dans ce cas, le terme de la méthode est inclus dans l'annotation de l'entité. Les cas les plus complexes sont les expressions telles que la valeur de sédimentation de Zeleny (*Zeleny sedimentation value*) qui sont le plus souvent utilisées à la place du nom du caractère, le volume de sédimentation de la farine (*flour sedimentation volume*) lié à la teneur en protéines du grain (*grain protein content*). Il a été décidé de les annoter en tant que traits.
L'annotation d'articles récents dans le corpus a révélé l'évolution du domaine en raison des progrès des méthodes de mesure. De nouveaux traits sont étudiés et de nouveaux termes apparaissent pour les désigner. Des exemples de ces nouveaux sujets sont liés à l'architecture des plantes (talles et racines), à la physiologie (réponse hormonale, relation source-puits) et à la reproduction. Nous avons ajouté 67 nouvelles classes spécifiques à WTO pour refléter l'évolution du domaine. La nouvelle version est disponible à l'adresse suivante : http://agroportal.lirmm.fr/ontologies/WHEATPHENOTYPE.
Le contrôle de cohérence révèle des erreurs, dont les plus fréquentes sont des annotations de classes trop spécifiques ou trop générales pour des expressions peu fréquentes, car la taille de WTO est trop importante pour que les experts puissent toutes les mémoriser. Ces erreurs ont été détectées automatiquement et corrigées manuellement.

Taxons

Les espèces sont également désignées par leur nom scientifique (par exemple *Triticum aestivum* L.), parfois abrégé (par exemple *T. aestivum*), et leur nom vernaculaire (par exemple blé tendre, orge, riz). Les noms vernaculaires peuvent être ambigus, en particulier le blé tendre qui est souvent appelé blé (*wheat*), bien que le blé soit le nom vernaculaire général du genre *Triticum* qui comprend 19 espèces, et pas seulement l'espèce *Triticum aestivum*.

Distribution de *Taec*

TaeC (Nédellec et al. 2023) est disponible sous licence CC-BY-ND à l'adresse https://entrepot.recherche.data.gouv.fr/dataset.xhtml?persistentId=doi:10.57745/GCYG3Q. Il est fourni dans le format d'annotation *standoff* BioNLP-ST (Bossy et al., 2013) qui est adapté à la représentation d'entités liées à une référence sémantique (Tableau 5). Le texte, les entités et les références sémantiques se trouvent dans des fichiers séparés. Les fichiers sont liés par le nom du fichier qui joue le rôle d'identifiant.

Table 5. File.txt contient le texte du document. File.a1 contient les annotations des entités nommées, leur identifiant, leur type et leur position. File.a2 contient les annotations sémantiques, le nom de la référence et l'identifiant de la référence.

File.txt
Efficiently tracking selection in a multiparental population: the case of earliness in wheat.

File.a1

T1 Trait	75	83	earliness
T2 Species	88	92	wheat

File.a2

N1	NCBI_Taxonomy Annotation:T1 Referent: 4565
N3	WTO Annotation:T2 Referent:WTO: 0000100

Le corpus de 540 documents est destiné à l'entraînement et à l'évaluation d'outils de prédiction et à leur comparaison. Dans ce dessein nous avons divisé l'ensemble de documents en trois sous-ensembles, Entraînement, Développement et Test. Nous distribuons publiquement les deux premiers sous-ensembles Entraînement et Développement. Nous organiserons en 2024 une compétition avec les données de Test dont les annotations restent confidentielles. La proportion des différentes parties est la suivante, 4/9, 1/9 et 1/3, soient 238, 116 et 174. Sur le total de 528 documents, sont donc distribués 360 documents.
La version de de WTO est la version 2.3.
(http://agroportal.lirmm.fr/ontologies/WHEATPHENOTYPE)

Conclusion

Dans cet article, nous avons présenté les motivations de la construction du nouveau corpus de référence *Taec*, composé de 540 documents relatifs à la recherche sur la sélection du blé et, plus généralement, au phénotypage des plantes cultivées. *Taec* est annoté avec trois types d'entités, à savoir les traits, les phénotypes et les espèces, et chaque entité est liée à une classe de référence de la taxonomie *NCBI* pour les espèces et de *WTO* pour les traits et les phénotypes, ce qui donne 10 635 occurrences de 340 classes.
Dans le cadre de travaux futurs à long terme, nous envisagerons l'annotation des variétés, des gènes et des marqueurs afin de répondre au besoin d'un corpus de relations génotypiques-phénotypiques pour la sélection végétale assistée par marqueurs.
À court terme, nous entraînerons des méthodes de REN/LEN avec *Taec* pour les utiliser dans le projet *D2KAB* pour annoter automatiquement toutes les références *PubMed* sur le blé. La base de connaissances sur le blé en cours de développement dans le projet *D2KAB* combinera ces annotations textuelles et les annotations du système d'information sur le blé *WheatIS* indexées par l'ontologie *CO_321* (http://www.wheatis.org/). L'alignement des classes de *WTO* et de l'ontologie *CO_321* permettra à l'utilisateur d'interroger les données d'observation et les documents de manière transparente et de tirer pleinement parti de la complémentarité des deux sources.

Remerciements

Les auteurs remercient Léonard Zweigenbaum (INRAE) pour sa contribution à l'annotation et Thierry Marcel et Jacques Le Gouis de l'INRAE pour leur contribution aux questions scientifiques relatives à la description du phénotypage du blé.
Les auteurs remercient la plateforme Migale pour avoir fourni les ressources nécessaires à l'exécution des services AlvisNLP (MIGALE, INRAE, 2020. (MIGALE, INRAE, 2020.
Migale Bioinformatics Facility, doi: 10.15454/1.5572390655343293E12).
Ce travail a été réalisé avec le soutien du projet *Data to Knowledge in Agronomy and Biodiversity* (D2KAB - www.d2kab.org) qui a bénéficié d'un financement de l'Agence Nationale de la Recherche (ANR-18-CE23-0017).

Références

Robert Bossy, Louise Deléger, Estelle Chaix, Mouhamadou Ba, Claire Nédellec. 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, nov 4, 2019 Association for Computational Linguistics 2019, ISBN 978-1-950737-82-6. doi 10.18653/v1/D19-5719 https://www.aclweb.org/anthology/D19-5719/

Estelle Chaix, Bertrand Dubreucq, Abdelhak Fatihi, Dialekti Valsamou, Robert Bossy, Mouhamadou Ba, Louise Deléger, Pierre Zweigenbaum, Philippe Bessières, Loïc Lepiniec, Claire Nédellec. Overview of the Regulatory Network of Plant Seed Development (SeeDev) Task at the BioNLP Shared Task. In *Proceedings of the BioNLP Shared Task 2016 Workshop*, Association for Computational Linguistics, Berlin, Germany 2016. 10.18653/v1/W16-3001

Cooper, L., Meier, A., Laporte, M. A., Elser, J. L., Mungall, C., Sinn, B. T., ... & Jaiswal, P. (2018). The Planteome database: an integrated resource for reference ontologies, plant genomics and phenomics. *Nucleic acids research*, *46*(D1), D1168-D1180.

Ćwiek-Kupczyńska, H., Altmann, T., Arend, D., Arnaud, E., Chen, D., Cornut, G., ... & Krajewski, P. (2016). Measures for interoperability of phenotypic data: minimum information requirements and formatting. *Plant Methods*, *12*, 1-18.

Gremse, M., Chang, A., Schomburg, I., Grote, A., Scheer, M., Ebeling, C., & Schomburg, D. (2010). The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources. *Nucleic acids research*, *39*(suppl_1), D507-D513.

Haq, H. U., Kocaman, V., & Talby, D. (2021). Deeper clinical document understanding using relation extraction. *arXiv preprint arXiv:2112.13259*.

Köhler S, Carmody L, Vasilevsky N, Jacobsen JOB, Danis D, Gourdine JP, Gargano M, Harris NL, Matentzoglu N, … Haendel MA, Mungall C, Robinson PN. Expansion of the Human Phenotype Ontology (HPO) knowledge base and resources. *Nucleic Acids Res.* 2019 Jan 8;47(D1):D1018-D1027. doi: 10.1093/nar/gky1105. PMID: 30476213; PMCID: PMC6324074.

Liang, C., Jaiswal, P., Hebbard, C., Avraham, S., Buckler, E. S., Casstevens, T., ... & Stein, L. (2007). Gramene: a growing plant comparative genomics resource. *Nucleic Acids Research*, *36*(suppl_1), D947-D953.

Lobo M, Lamurias A, Couto FM. Identifying Human Phenotype Terms by Combining Machine Learning and Validation Rules. *Biomed Res Int.* 2017;2017:8565739. doi: 10.1155/2017/8565739. Epub 2017 Nov 9. PMID: 29250549; PMCID: PMC5700471.

Nédellec, Claire, Clara Sauvion, Louise Deléger, Robert Bossy, et Léonard Zweigenbaum. « Triticum aestivum trait Corpus ». Recherche Data Gouv, 2023. https://doi.org/10.57745/GCYG3Q.

Claire Nédellec, Robert Bossy, Dialekti Valsamou, Marion Ranoux, Wiktoria Golik, Pierre Sourdille. Information Extraction from Bibliography for Marker Assisted Selection in Wheat. In *proceedings of Metadata and Semantics for Agriculture, Food & Environment (AgroSEM'14), special track of the 8th Metadata and Semantics Research Conference (MTSR'14),* Springer Communications in Computer and Information Science, Series Volume 478, Karlsruhe, pp 301-313, Germany, 2014. DOI: 10.1007/978-3-319-13674-5_28.

Nédellec C., Bossy R., Chaix E., Deléger L. Text-mining and ontologies: new approaches to knowledge discovery of microbial diversity. Article et conférence invitée. In *Proceedings of the 4th International Microbial Diversity Conference*. pp. 221-227, ed. Marco Gobetti. Bari, Pub. Simtra. ISBN 978-88-943010-0-7, Bari, Italy, October 2017. arXiv:1805.04107

Claire Nédellec, Liliana Ibanescu, Robert Bossy, Pierre Sourdille. WTO, an ontology for wheat traits and phenotypes in scientific publications. 18(2) *Genomics & Informatics*. June 2020. doi: 10.5808/GI.2020.18.2.e14

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.

Paux, E., Lafarge, S., Balfourier, F., Derory, J., Charmet, G., Alaux, M., ... & Breedwheat Consortium. (2022). Breeding for Economically and Environmentally Sustainable Wheat Varieties: An Integrated Approach from Genomics to Selection. *Biology*, *11*(1), 149.

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., ... & Lewis, S. (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology*, *25*(11), 1251-1255.

Sousa, D., Lamurias, A., Couto, F.M.: A silver standard corpus of human phenotype-gene relations. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 1487–1492. Association for Computational Linguistics, Minneapolis (2019)

Dóra Szakonyi, Sofie Van Landeghem, Katja Baerenfaller, Lieven Baeyens, Jonas Blomme, Rubén Casanova-Sáez, Stefanie De Bodt, …Wilhelm Gruissem, Sean Walsh, Pierre Hilson, The KnownLeaf literature curation system captures knowledge about Arabidopsis leaf growth and development and facilitates integrated data mining, *Current Plant Biology*, Volume 2, May 2015, Pages 1-11, ISSN 2214-6628, http://dx.doi.org/10.1016/j.cpb.2014.12.002.

Sevgili, Ö., Shelmanov, A., Arkhipov, M., Panchenko, A., & Biemann, C. (2022). Neural entity linking: A survey of models based on deep learning. *Semantic Web*, (Preprint), 1-44.

Tadesse, W., Sanchez-Garcia, M., Assefa, S. G., Amri, A., Bishaw, Z., Ogbonnaya, F. C., & Baum, M. (2019). Genetic gains in wheat breeding and its role in feeding the world. *Crop Breed. Genet. Genom*, *1*, e190005.

Van Landeghem S, De Bodt S, Drebert ZJ, Inzé D, Van de Peer Y. The potential of text mining in data integration and network biology for plant research: a case study on Arabidopsis. *Plant Cell*. 2013 Mar;25(3):794-807. doi: 10.1105/tpc.112.108753.

Wilkinson, M. D., Dumontier, M., Aalbersberg, I. J., Appleton, G., Axton, M., Baak, A., ... & Mons, B. (2016). The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, *3*(1), 1-9.