



Wheat Trait and Phenotype : une ontologie pour l'étude et l'annotation sémantique de données et de textes.

Les ontologies sont au cœur du développement du web sémantique afin de standardiser les échanges de données entre machines distantes. Dans le cadre du projet ANR D2KAB, plusieurs équipes INRAE de disciplines différentes ont collaboré au développement de l'ontologie Wheat Trait and Phenotype Ontology (WTO) pour définir les classes de facteurs et les traits phénotypiques associés au blé tendre.



CONTEXTE ET ENJEUX

Les variétés de blé présentent une grande diversité de traits et de phénotypes. Les expériences de sélection variétale en lien avec le génotype sont documentées par un grand nombre d'ouvrages scientifiques et de données d'observation. Le recoupement de ces informations complémentaires est essentiel à l'étude de la relation génotype-phénotype et, plus généralement, à l'amélioration de la sélection du blé pour répondre notamment aux enjeux de changement climatique et d'alimentation.

Le portail de découverte de données international Faidare (FAIR Data-finder for Agronomic Research) a pour ambition d'agréger ces informations et d'offrir aux utilisateurs une interface pour exprimer des requêtes fédérées de manière expressive et transparente grâce à la standardisation et à la structuration du vocabulaire de traits et phénotype. Faidare utilise la Crop ontology CO_321 pour indexer les informations de terrain avec leurs protocoles et unités. Nous avons développé l'ontologie Wheat Trait and Phenotype Ontology (WTO) comme moyen adapté à l'extraction et l'indexation d'information de la littérature scientifique par des méthodes de traitement automatique de la langue (TAL) qui soit interopérable avec Crop ontology CO_321.



RESULTATS

WTO est le résultat d'un travail interdisciplinaire de longue haleine. Elle décrit les facteurs environnementaux et les traits phénotypiques dans toute leur diversité en 748 classes : développement, agronomie, résistance au stress, qualité boulangère, etc.. Les classes de résistances aux maladies fongiques sont notamment très détaillées. L'abondance des classes et des synonymes en fait une ressource sémantique complète dont la navigation est facilitée par sa structure profonde.

Notre workflow d'extraction d'information de la littérature basé sur AlvisNLP extrait et normalise les mentions de traits phénotypiques par les classes de WTO, interrogeables par le moteur de recherche bibliographique AlvisIRSamBlé opéré par la plateforme Migale.

Pour évaluer la qualité de l'extraction, nous avons créé le corpus Taec (*Triticum aestivum* corpus) annoté manuellement par les mentions de taxon et de traits phénotypiques de 520 documents suivant des consignes détaillées

L'intégration de données expérimentales et textuelles dans Faidare a nécessité l'alignement complexe des classes des deux ontologies WTO et CO_321 par un travail interdisciplinaire étroit. L'alignement est représenté dans le standard SSSOM (Simple Standard for Sharing Ontological Mappings)- standard à l'évolution ce travail a contribué – et justifié par la formalisation de règles explicites d'alignement. WheatGenomicsSLKG est un premier prototype intégré réalisé en RDF grâce à la conversion de WTO en SKOS/OWL

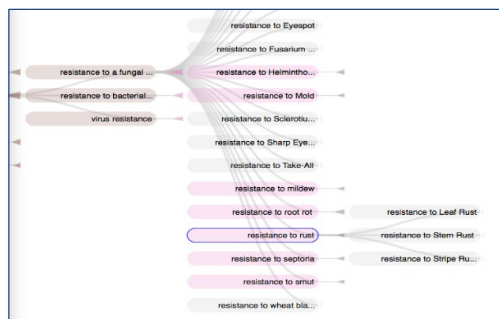


PERSPECTIVES

Nos projets de partenariat avec l'URGI et PlantBioinfoPF compléteront nos résultats par un workflow d'annotation sémantique des allèles/gènes et cultivars pour le blé tendre et l'élargiront à d'autres espèces cultivées.

L'ontologie WTO et le corpus TaeC seront proposés pour l'organisation d'une shared task en normalisation d'entité (ou entity-linking) à CLEF (Conference and Labs of the Evaluation Forum) en 2025 avec l'objectif de comparer les performances de méthodes de TAL.

L'ontologie WTO, son alignement avec CO_321 et le corpus TaeC sont publiés sur le dépôt RDG. Ces résultats ont été obtenus dans le projet ANR ANR-18-CE23-0017 D2KAB (Data to Knowledge in Agriculture and Biodiversity). L'intégration et la mise en production dans le portail Faidare sont programmées à court terme. WTO a inspiré la structure de l'ontologie des traits phénotypique du riz dans le cadre d'une collaboration avec l'Université d'Agriculture Huazhong, Wuhan, qui se poursuit.



Figures : Un exemple de recherche d'information utilisant WTO pour identifier les variétés et leur génotype en lien avec un trait de résistance aux bioagresseurs.



REFERENCES

<http://doi.org/10.15454/1.4382637738008071E12>

Données

- > WTO AgroPortal: <https://tinyurl.com/4bcksm84>
- > AlvisIR : <https://bibliome.migale.inrae.fr/wheat/alvisir/webapi/search>
- > Triticum aestivum trait Corpus, (Taec) : <https://doi.org/10.57745/GCYG30>
- > Alignment of WTO and CO_321 ontology classes : <https://doi.org/10.57745/ZLJYQ0>
- > Nédellec C., Sauvion C., Bossy R., Borovikova M., Deléger L. (2024) - TaeC: a Manually annotated text dataset for trait and phenotype extraction and entity linking in wheat breeding literature. Plos One. June 2024. 10.1371/journal.pone.0305475
- > Nédellec C., Aubin S., Sauvion C., Ibanescu L., Bravo S., Le Gouis J., Marcel T., Pommier C., Bossy R., Alaux M. (2024) - Mapping bread wheat trait ontologies for semantic interoperability. F1000Research. <https://doi.org/10.12688/f1000research.154860.1>
- > Yacoubi Ayadi N., Bernard S., Bossy R., Courtin M., Gates Happi Happi B., Larmande P., Michel F., Nédellec C., Roussey C., Faron C. (2024) - A Unified Approach to Publish Semantic Annotations of Agricultural Documents as Knowledge Graphs. Smart Agricultural Technology, 2024.100484, ISSN 2772-3755, <https://doi.org/10.1016/j.atech.2024.100484>.



Contact : Claire Nédellec

Unité : MaIAGE

Département : MathNum

Centre INRAE : Île de France Jouy-en-Josas

INRAE 2030 Orientations scientifiques (OS) et Orientations de politique générale (OP) : OS 5.1, OS 1.3, OS 2.2, OP 1.3

Grands Objectifs Scientifiques (GOS) MathNum : GOS 1

Mots-clés : Ontologie, blé, trait, sélection variétale, traitement de la langue, intégration de données, alignement d'ontologie, graphe de connaissance