Aligner les descriptions des plantes ayant des points de vue distincts

F. Amardeilh[1], S. Aubin[2], S. Bernard[3], S. Bravo[2], R. Bossy [4], C. Faron[5], F. Michel[5], C. Roussey[6]

[1] Elzeard, Bordeaux, France
[2] DIPSO, INRAE, France
[3] LISC, INRAE, Aubière, France
[4] MaIAGE, INRAE, Jouy-en-Josas, France
[5] Université Côte d'Azur, INRIA, CNRS, I3S, France
[6] MISTEA, INRAE, Montpellier, France

florence.amardeilh@elzeard.co, sophie.aubin@inrae.fr, stephan.bernard@inrae.fr, robert.bossy@inare.fr, faron@i3s.unice.fr, fmichel@i3s.unice.fr, catherine.roussey@inrae.fr

1 Contexte

Nous présentons nos travaux sur l'alignement de deux graphes de connaissances complémentaires utiles dans le domaine agricole : le thésaurus des Usages des plantes Cultivées en France (FCU) et le REFérentiel TAXonomique national français TAXREF pour la faune, la flore et les champignons déjà publié dans [1]. FCU décrit l'utilisation des plantes en agriculture, ou plus exactement le rôle des plantes en agriculture : par exemple les '*tomates*' sont des cultures utilisées pour l'alimentation humaine. Il représente le point de vue des agriculteurs. TAXREF décrit les taxons biologiques et les noms scientifiques associés, ou plus exactement une description de cette plante suivant sa composition génétique : par exemple, une espèce de tomate peut être '*Solanum lycopersicum*'. TAXREF est maintenu par le Muséum national d'Histoire naturelle (MNHN) et représente le point de vue des taxonomistes et des agronomes. Les deux graphes de connaissances contiennent des noms vernaculaires de plantes. Les noms vernaculaires sont souvent ambigus et peu consensuels, ce qui rend l'activité d'alignement particulièrement difficile.

2 Travaux antérieurs

Nos travaux précédents [4] ont implémenté plusieurs méthodes d'alignement automatique basées sur la comparaison de noms vernaculaires. Ces méthodes automatiques ont réutilisé des sources de référence existantes comme la base de données européenne de l'EPPO [1] et le catalogue officiel français des espèces et variétés de plantes cultivées du GEVES [2]. Les résultats montrent qu'il est nécessaire de nettoyer les alignements produits automatiquement en raison de l'ambiguïté des noms vernaculaires. Par conséquent, un groupe d'experts agricoles a produit des alignements valides.

1. https ://gd.eppo.int/
2. Catalogue officiel des esp 'eces et variétés de plantes cultivéesées en France accessible sur https ://www.geves.fr/catalogue/

3 Matériels

3.1 TAXREF et TAXREF-LD

TAXREF [3] est le référentiel taxonomique français de la faune, de la flore et des champignons. TAXREF est disponible sous la forme d'un graphe de connaissances respectant les principes des données liées, nommé TAXREF-LD [5]. TAXREF-LD est disponible sur AgroPortal [3]. Ce travail a été développé en utilisant la version 15.2 de TAXREF-LD qui contient 287 229 classes et plus de 1 000 000 instances.

3.2 Thésaurus FCU

Le thésaurus des Usages des plantes Cultivées en France normalise les noms de cultures en français. De plus, il organise ces noms de cultures en catégories selon leurs usages sur le territoire français. Les usages représentent également les secteurs agricoles. Le thésaurus est publié sur le Web selon les principes des données liées. Le thésaurus est disponible sur AgroPortal [4]. Ce travail a été développé en utilisant la version 3.3 de FCU qui contient 707 instances de `skos:Concept`.

3.3 Propriétés d'alignement

Dans TAXREF-LD, un taxon est défini par une classe `owl:Class` et les noms de taxon sont définis par des instances de `skos:Concept`. Une classe regroupe l'ensemble des spécimens caractérisant actuellement le taxon considéré. Ces spécimens sont généralement stockés par les muséums d'histoire naturelle ou par les centres de ressources génétiques. Les noms scientifiques découlent des avancées de la taxonomie : la science de classer et nommer les êtres vivants. Les méthodes de classification, de création des taxons, ont évoluées avec le temps. Ainsi, les noms scientifiques d'espèces et leurs synonymes sont une trace de ces évolutions successives des taxons. Dans FCU, un rôle de plante cultivée est défini

3. https ://agroportal.lirmm.fr/ontologies/TAXREF-LD
4. https ://agroportal.lirmm.fr/ontologies/CROPUSAGE

par une instance de `skos:Concept`. Un concept FCU a pour label un nom vernaculaire enrichi de son rôle : par exemple le concept "carotte potagère" représente les carottes consommées pour l'alimentation humaine. Nous avons défini 12 propriétés d'objet pour lier une instance de `skos:Concept` représentant un nom scientifique à une instance de `skos:Concept` de FCU représentant un rôle de plante en agriculture. Ces propriétés permettent de répondre à la question de compétence : "quel nom scientifique (principalement de rang espèce) peut jouer ce rôle en agriculture ?". Sachant qu'il est possible qu'un nom scientifique d'espèce ne réponde pas entièrement au rôle, et inversement. Ces propriétés d'objets peuvent être vu comme une spécialisation de la propriété `skos:closeMatch` : un nom scientifique identifiant une espèce n'est pas équivalent à un rôle agricole mais il est proche. Par exemple, toutes les espèces de carottes cultivées ont potentiellement deux rôles en agriculture : carottes potagères (alimentation humaine) et carottes fourragères (alimentation animale). Ce qui n'est pas le cas pour la chicorée qui a des espèces différentes remplissant l'un des deux usages agricoles : chicorées potagères et chicorées industrielles. Pour compléter, nous avons aussi défini 10 propriétés d'annotations pour lier une `owl:Class` représentant un taxon à une instance de `skos:Concept` de FCU représentant un rôle de plante en agriculture. Ainsi, les triplets utilisant les propriétés d'objet sont documentés (annotés) par les triplets utilisant les propriétés d'annotation.
Il n'existe pas de règles d'alignement entre les espèces (noms scientifiques) et les rôles agricoles, ce qui explique le besoin d'aligner manuellement ces différentes entités. L'ensemble de ces nouvelles propriétés sont disponible dans une ontologie French Crop Usage Ontology disponible sur AgroPortal [5].

3.4 Schéma d'alignement SSSOM

"Simple Standard for Sharing Ontology Mappings" (SSSOM) est un standard récent, développé par la communauté biomédicale autour de OBO Foundry et décrit dans [2]. Pour ce travail, nous avons utilisé la version 0.11.O de SSSOM sortie en mars 2023.

4 Méthode d'alignement manuelle

Suite aux résultats non satisfaisants des méthodes automatiques, nous avons créé un nouvel ensemble d'alignements en demandant à des experts de proposer des alignements corrects et réputés entre des rôles de plantes en agriculture, les taxons et leurs noms scientifiques. Ainsi, tous les alignements proposés doivent avoir une valeur de confiance élevée. En cas d'ambiguïté, l'alignement ne doit pas être créé. Dans un premier temps, nous avons fourni aux experts des consignes pour les aider dans leurs décisions. Deuxièmement, quatres outils de recherche ont été proposés pour rechercher des termes dans les deux graphes de connaissances : AgroPortal, deux SPARQL end points et un outil de recherche par facettes. Troisièmement, trois règles de conservation ont été écrites pour contextualiser les alignements créés et indiquer leur provenance. Nous nous sommes concentrés sur des cultures spécifiques : vigne, carotte, chicorée et tomate en fonction de la disponibilité des experts.

5 Résultats

Les alignements réalisés par les experts en suivant la méthode précédente sont publiés en données ouvertes sur le référentiel français Recherche Data Gouv [6]. Ainsi, ils pourraient être utilisés comme base de référence pour valider toute méthode d'alignement automatique.

Références

[1] Florence Amardeilh, Sophie Aubin, Stephan Bernard, Catherine Faron, Sonia Bravo, Robert Bossy, Franck Michel, Juliette Raphel, and Catherine Roussey. Combining different points of view on plant descriptions : mapping agricultural plant roles and biological taxa. *Frontiers in Artificial Intelligence*, 6 :118803, 2023.

[2] Matentzoglu et al. A Simple Standard for Sharing Ontological Mappings (SSSOM). *Database*, 2022, 05 2022. baac035.

[3] Olivier Gargominy, Sandrine Tercerie, C Régnier, T Ramage, P Dupont, P Daszkiewicz, and L Poncet. TAXREF v15, référentiel taxonomique pour la France : méthodologie, mise en œuvre et diffusion. Technical report, 2021.

[4] Franck Michel, Florence Amardeilh, Robert Bossy, Catherine Faron, Catherine Roussey, and Camille Noûs. Alignement entre sources : cas d'usage des plantes cultivées. In *Journées francophones d'Ingénierie des Connaissances*, Saint-Étienne, France, June 2022.

[5] Franck Michel, Olivier Gargominy, Sandrine Tercerie, and Catherine Faron-Zucker. A Model to Represent Nomenclatural and Taxonomic Information as Linked Data. Application to the French Taxonomic Register, TAXREF. In *Proceedings of the ISWC2017 workshop on Semantics for Biodiversity (S4BioDiv)*, volume 1933, Vienna, Austria, 2017. CEUR Workshop Proceedings. http ://ceur-ws.org/Vol-1933/paper-3.pdf.

Remerciements

Nous tenons à remercier les experts agronomes : Thierry Lacombe (orcid :0000-0001-9968-8228), Olivier Yobregat (orcid :0000-0002-7516-8727) et Juliette Raphel (orcid :0000-0002-5872-5034). Ces travaux ont été financé par projet ANR Data to Knowledge in Agronomy and Biodiversity (ANR-18-CE23-0017) et par le Plan de Relance et le Programme d'Investissements d'Avenir «i-Nov» du gouvernement français.

5. https ://agroportal.lirmm.fr/ontologies/FCUO

6. https ://doi.org/10.57745/LVRFWJ