Annotation des stades phénologiques et cultures dans les Bulletins de Santé du Végétal de la vigne

Marine Courtin, Stephan Bernard, Robert Bossy, Catherine Roussey

Annotation des stades phénologiques et cultures dans les Bulletins de Santé du Végétal de la vigne

M. Courtin[1,2], S. Bernard[1], R. Bossy[2], C. Roussey[1,3]

[1] Université Clermont Auvergne, INRAE, UR TSCF, Aubière, France
[2] Université Paris-Saclay, INRAE, UR MAIAGE, 78350 Jouy-en-Josas, France
[3] INRAE, UR MISTEA, Montpellier, France

marine.courtin@inrae.fr, stephan.bernard@inrae.fr, robert.bossy@inrae.fr, catherine.roussey@inrae.fr

Résumé

Dans cet article, nous présentons une application des méthodes de reconnaissance et de normalisation d'entités nommées pour l'annotation sémantique des cultures et stades phénologiques dans les Bulletins de Santé du Végétal. La méthode proposée utilise des ressources sémantiques préexistantes : FrenchCropUsage pour les cultures et PPDO pour les stades phénologiques, et introduit une désambiguisation des entités basée sur la structure de ces graphes de connaissances. À terme, cette annotation pourra être exploitée pour proposer aux conseillers agricoles et aux acteurs de la biovigilance en agriculture une interface de lecture sémantiquement interrogeable des Bulletins de Santé du Végétal.

Mots-clés

Agriculture, Graphes de connaissances, Modélisation Sémantique, Traitement Automatique de la Langue, Annotation, Ressources Sémantiques, Reconnaissance d'Entités Nommées, Normalisation d'Entités Nommées.

Abstract

In this paper, we apply named entity recognition and normalization methods to the semantic annotation of crops and phenological stages in Plant Health Bulletins. The proposed method uses pre-existing semantic resources : FrenchCropUsage for crops and PPDO for phenological stages, and introduces entity disambiguation based on the structure of these knowledge graphs. Eventually, this annotation could be used to offer agricultural advisers and biovigilance players in agriculture a semantically queriable reading interface for Plant Health Bulletins.

Keywords

Agriculture, Knowledge Graphs, Semantic modelling, Natural Language Processing, Annotations, Semantic Resources, Entity Recognition and Linking.

1 Introduction

Les acteurs du monde agricole sont confrontés à des enjeux importants du point de vue économique, social et écologique. Pour faire face à ces enjeux, en comprendre les mécanismes et y trouver des réponses, l'accès à l'information tient un rôle capital. Les techniques de TAL (Traitement Automatique de la Langue), d'Ingénierie des Connaissances et du Web Sémantique permettent d'extraire et synthétiser l'information contenue dans les données, de la représenter et de la rendre disponible et interopérable afin d'en faciliter l'accès. Nous présentons ici notre travail visant à enrichir les Bulletins de Santé du Végétal au travers d'une annotation sémantique sur certaines des entités clés qu'ils contiennent : les cultures et les stades phénologiques. La structure de l'article est la suivante. Nous commençons par présenter les matériels, à savoir le corpus des Bulletins de Santé du Végétal et les ressources sémantiques utilisées, ainsi que les pré-traitements qui leur sont appliqués. Puis, nous décrivons nos méthodes pour l'identification et la normalisation des entités de type culture et de stade phénologique dans ce corpus. Ces entités présentent des ambiguïtés que nous introduisons brièvement avant de décrire deux méthodes de désambiguïsation. Nous élargissons ensuite ce travail en le reliant à d'autres travaux dont les problématiques sont similaires, avant de conclure et de décrire quelques perspectives, notamment sur l'évaluation de notre travail.

2 Matériels

2.1 Corpus des Bulletins de Santé du Végétal

En France, le Grenelle de l'environnement et le plan Ecophyto ont renforcé les réseaux nationaux de surveillance sur les cultures et les pratiques agricoles. Les Bulletins de Santé du Végétal sont une des modalités mises en place par ces réseaux de surveillance dans l'ensemble des régions et départements d'outre-mer. Le Bulletin de Santé du Végétal (BSV) est un document d'information à la fois technique et réglementaire, rédigé sous la responsabilité d'un comité régional d'épidémiosurveillance. Un BSV a pour objectif de réunir et présenter les actualités majeures concernant l'état sanitaire d'un ensemble de cultures sur une région donnée. Il rapporte des observations sur le développement des cultures, les attaques de bioagresseurs, et présente une analyse du risque phytosanitaire dans la région. Près de 15 000 parcelles sont observées chaque année pour éditer en-

viron 3400 BSV par an. Depuis le début de leur parution, les BSV sont librement accessibles au format PDF sur différents sites internet. Le laboratoire TSCF a collecté 36469 Bulletins de 2009 à 2022 concernant l'ensemble du territoire français [6].
Le corpus des BSV utilisé dans nos travaux est constitué de 77 fichiers concernant la vigne provenant du corpus de test D2KAB. Ces BSV ont été publiés en 2018 et 2019 et concerne toutes les régions viticoles [1]. Les BSV sont convertis en HTML à l'aide d'un outil nommé `pdf2blocks` [2] qui s'efforce de reconstruire la structure des titres et sous-titres du bulletin.

2.2 Ontologie PPDO et son graphe de connaissances

L'échelle phénologique BBCH améliorée (Biologische Bundesanstalt, Bundessortenamt und CHemische Industrie) propose une codification homogène des stades de développement communs à différentes espèces végétales cultivées [3]. BBCH décrit plusieurs ensembles de stades de développement : l'échelle générale et des échelles spécifiques par culture (dites « échelles individuelles »). Il existe par exemple une échelle individuelle pour la vigne.
BBCH-based Plant Phenological Description Ontology (PPDO) [6] est une extension du vocabulaire SKOS [4] pour décrire les échelles phénologiques des plantes fondée sur la codification BBCH. Une échelle phénologique est une instance de $skos{:}ConceptScheme$ et un stade est une instance de $skos{:}Concept$. Dans [6] nous avons décrits et alignés cinq échelles phénologiques de la vigne en utilisant les propriétés comme $skos{:}exactMatch$ et $skos{:}closeMatch$. La figure 1 présente un alignement entre un stade de l'échelle BBCH individuelle de la vigne et un stade d'une autre échelle phénologique de la vigne intitulée IFV label. L'ontologie est publiée sur Agroportal. Le graphe de connaissances composé de l'ensemble des échelles et de leur stades est disponible sur un répertoire git [3] et est interrogeable sur un SPARQL EndPoint [4].

2.3 FCU

Le thésaurus intitulé "usages des plantes cultivées en France" ou French Crop Usage (FCU) [5] normalise les noms de plantes cultivées en français. De plus, il les organise dans des catégories représentant des filières agricoles : par exemple, *"fourrage"* et *"grandes cultures"* sont deux exemples de filières agricoles. Ainsi, une hiérarchie est formée par des relations de généralisation/spécialisation entre les filières agricoles et les noms d'usage des plantes cultivées : par exemple, *"grandes cultures"* se spécialise en *"céréales"*. Les termes du thésaurus ont été sélectionnés manuellement à partir de documents de référence : Les statistiques agricoles annuelles de l'Agreste [6], les métadonnées du registre parcellaire graphique, le classement des plantes cultivées par groupe d'usage proposé par wikipédia France, le catalogue officiel des espèces et variétés de plantes cultivées en France du GEVES, les fiches "les plantes fourragères pour les prairies" du GNIS [7], la base Ephy qui décrit l'usage des produits phytosanitaires sur les plantes, le Larousse Agricole.
Le choix des noms d'usage des plantes cultivées, les définitions associées et leur organisation ont été discutés par au moins un expert de la filière agricole. Ce thésaurus n'est pas complet et évolue en fonction des projets. Le thésaurus est modélisé à l'aide du vocabulaire SKOS proposé par le W3C [4], la figure 2 en présente un extrait. Il est disponible sur le Web de données liées [8]. FCU contient 526 $skos{:}Concept$. La profondeur maximale de la hiérarchie est de 6. Chaque concept est défini par un ensemble d'étiquettes (les noms vernaculaires de la plante), des notes, des liens vers d'autres sources d'information et de liens hiérarchiques.
Lorsqu'une plante a plusieurs usages, elle est représentée par plusieurs concepts : un concept pour chacun des usages, plus un concept pour l'ensemble des usages, parent des concepts précédents. Un concept dans la branche "multiusage" porte le nom vernaculaire de la plante sans indication d'usage (par exemple *"carotte"*). Ce concept est ensuite décliné en autant de fils qu'il y a d'usages (*"carotte potagère"* pour l'alimentation humaine et *"carotte fourragère"* pour l'alimentation animale). Chacun des fils est de plus positionné à un seul endroit dans la branche *"usage des plantes cultivées"*. Dans la figure 3, le concept *"carotte potagère"* est positionné comme fils du concept *"légume racine"*.

2.4 Pré-traitements

Une fois le corpus récupéré au format HTML, celui-ci est tokenisé, segmenté en phrases et lemmatisé par Stanza [5]. La même tokenisation et lemmatisation est appliquée sur PPDO et FCU, afin de disposer d'étiquettes ($skos{:}prefLabel$ et $skos{:}altLabel$) lemmatisées pour améliorer la projection des graphes de connaissances et neutraliser la variation morphologique.

3 Méthodes

3.1 Détection des stades phénologiques

Dans cette section nous décrivons les méthodes d'identification des mentions de cultures et stades phénologiques dans les BSV, ainsi que leur normalisation par les instances de FCU et PPDO. La chaîne de traitement mise en oeuvre utilise l'outil AlvisNLP [1].
Dans PPDO, les stades phénologiques qui sont des instances de *skos :Concept* sont associés à des étiquettes. Ces étiquettes, ainsi que leur version lemmatisée sont utilisées pour repérer les mentions de stades phénologiques dans le corpus. Chaque fois qu'il y a une correspondance entre la

1. les corpus de tests sont disponibles sur le répertoire git `https://forgemia.inra.fr/bsv/corpus-bsv/-/tree/master/corpus_test`
2. `https://doi.org/10.5281/zenodo.4067965`
3. `https://gitlab.irstea.fr/copain/phenologicalstages`
4. `http://ontology.inrae.fr/ppdo/sparql/`
5. `https://doi.org/10.15454/QHFTMX`
6. l'Agreste est le service statistique ministériel de l'agriculture
7. Le GNIS est l'interprofession des semences et plants, il a été renommé SEMAE
8. `http://ontology.irstea.fr/pmwiki.php/Site/FrenchCropUsage`

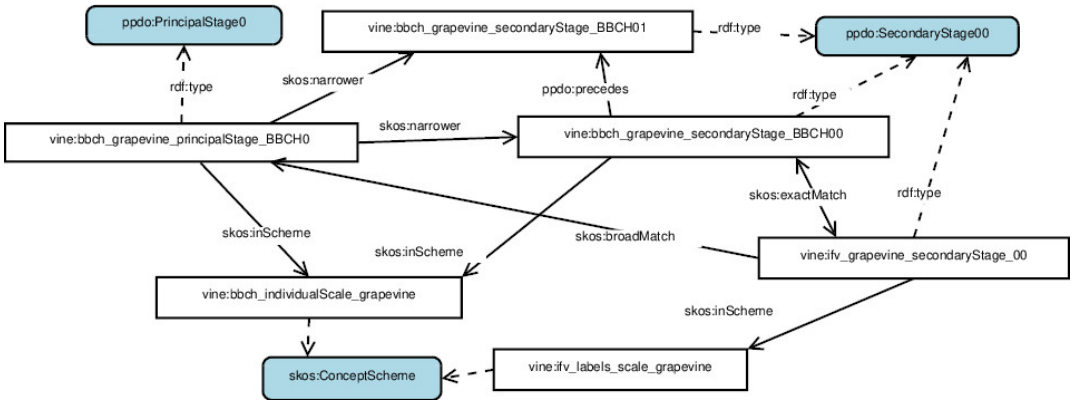

FIGURE 1 – extrait du graphe PPDO.

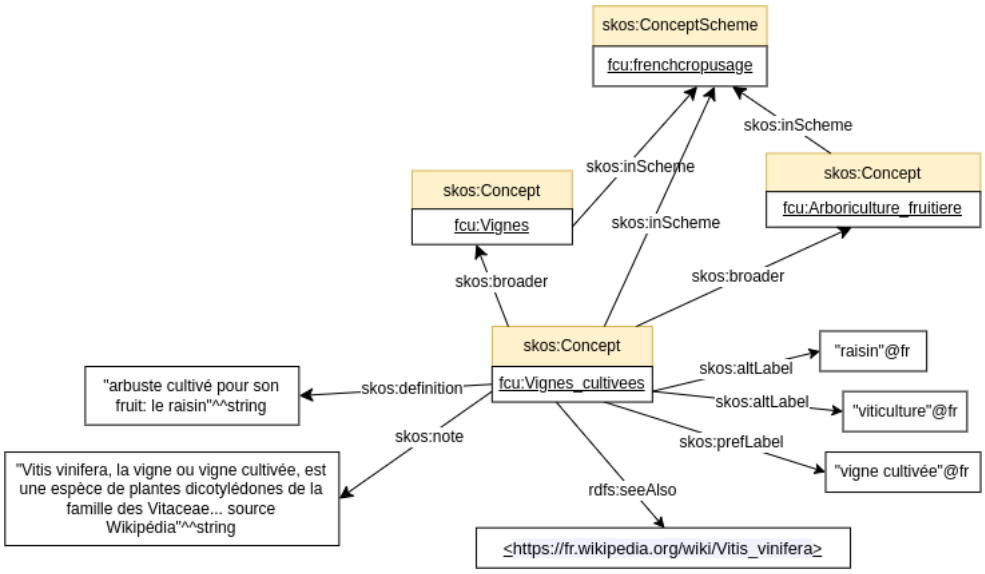

FIGURE 2 – Un extrait du thésaurus FCU présentant le concept de vigne cultivée.

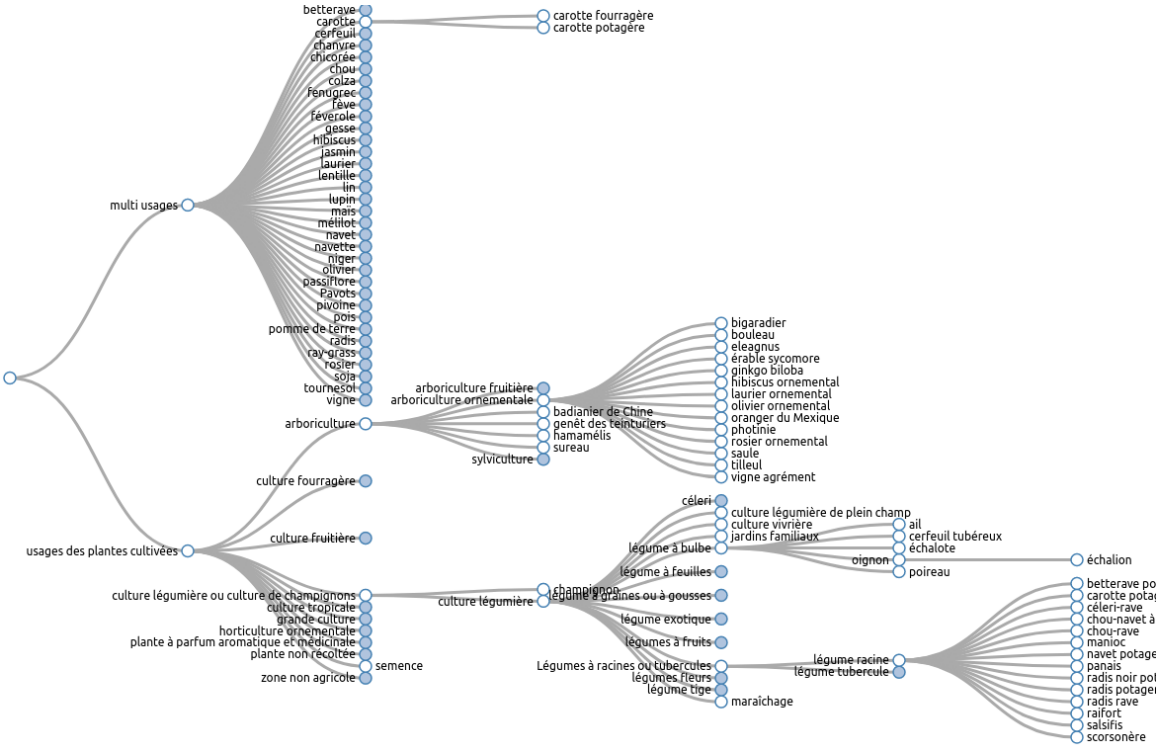

FIGURE 3 – Un exemple de visualisation du thésaurus avec l'outil SKOSPlay

forme ou les lemmes du texte et les étiquettes, nous créons une mention normalisée par l'uri associée au stade phénologique.

Lorsque les étiquettes des stades phénologiques ne permettent pas de détecter une mention car celle-ci est exprimée sous une forme différente, nous introduisons des motifs. Ainsi le stade IFV Epicure 36 est associé à deux étiquettes : "stade secondaire IFV-Epicure 36 de la vigne" et "Mi-véraison (50%) (M)". Cependant dans le texte, ce stade est exprimé sous d'autres formes, comme dans les phrases *«Le stade mi-véraison est atteint»* ou encore *«Ci-dessous, prévisions phénologiques pour le Riesling à Bergheim (Civa) : mi véraison entre le 12 et le 19/08»*. Un motif est introduit pour capturer ces formes : "mi véraison" et "mi-véraison".

```
<epicure-36 class="PatternMatcher">
    <pattern>
        ([str:lower(@form) == "mi"]
        [str:lower(@form)=="véraison"])
        |
        [str:lower(@form)=="mi-véraison"]
    </pattern>
    <actions>
       <createAnnotation layer="ifv-epicure" rank='36")'/>
    </actions>
  </epicure-36>
```

Les motifs permettent ainsi de reconnaître une partie des mentions, qui de part la variabilité de leur forme, ne correspondent pas exactement aux étiquettes du graphe de connaissances.

Une fois les mentions repérées, que ce soit par projection des étiquettes du graphe de connaissance, ou par l'application de motifs, nous disposons pour chaque mention de sa normalisation sous la forme du stade décrit dans PPDO. Nous utilisons alors les alignements entre les stades pour enrichir cette normalisation et attribuer à la mention tous les stades phénologiques qui lui correspondent. Ainsi pour la phrase *«Sur les parcelles à risque(régulièrement attaquées), les dégâts peuvent apparaître très précocement, dès le stade pointe verte.»*

La mention "pointe verte" est détectée et associée au stade Baggiolini C grâce à la correspondance entre le texte et l'étiquette de ce stade. Comme Baggiolini C est aligné avec 4 autres stades, la mention sera également normalisée par IFV Epicure 05, Eichhorn-Lorenz 05, BBCH (vigne) 07 et IFV Label 07. Lors de l'exploitation des annotations, quelle que soit l'échelle utilisée pour formuler la requête, cette mention sera retournée dans les résultats.

3.2 Traitement des ambiguïtés

Au cours de la reconnaissances des entités de type Culture et Stade Phénologique, certaines mentions sont introduites par erreur dans l'annotation. Ainsi on souhaiterait identifier que *"petits pois"* est un stade phénologique dans la phrase *«Le stade moyen actuel est petits pois»*, mais pas dans la phrase *«Nous avons observé des attaques de mildiou sur des petits pois en particuliers dans des exploitations en agriculture biologique»* (où il s'agit en réalité d'une autre entité, de type $Culture$). Nous présentons brièvement quelques cas d'ambiguïté communs pour les entités de type $Culture$ et $Stade\ Phnologique$.

Confusion sur les stades phénologiques Afin de caractériser les différents cas d'ambiguïté, nous commençons par

extraire tous les stades phénologiques de la vigne qui apparaissent dans des bulletins non-viticoles. Nous pouvons regrouper ces ambiguïtés en 4 catégories :

— la mention correspond à un stade phénologique, dans une échelle qui n'est pas spécifique à la vigne (*"floraison"*). Dans ce cas là, il s'agit bien d'un stade phénologique, et on ne considèrera pas que l'on a affaire à une erreur.
— la mention correspond à une culture (*"petits pois"*).
— la mention correspond au stade de développement d'un bioagresseur (*"maturité des larves"*).
— la mention correspond à un symptôme lié à la présence d'un bioagresseur (*"chute des feuilles"*).

Confusion sur les cultures Parmi les mentions qui sont erronément identifiées comme étant des cultures, on trouvera en autres des références à des bioagresseurs (*"champignon"*), des parties de plantes (*"baies"*), des symptômes liés à la présence d'un bioagresseur (*"bois noir"*), des descriptions météorologiques (*"gel"*[9], ou encore des noms de personnes ("véronique", "olivier").

Afin d'améliorer la précision des annotations, nous proposons d'introduire des scores de confiance qui permettrons de filtrer les annotations et d'enlever celles qui sont jugées peu fiables. Ces scores de confiance sont basés sur la connectivité des concepts repérés dans chaque bulletin et visent à favoriser les mentions présentant une connectivité jugée comme "désirable".

3.3 Scores de confiance

Hypothèse 1 : un stade phénologique est plus fiable si les stades proches sont également mentionnés dans le bulletin.

Puisque le bulletin couvre les observations sur une courte période (une semaine), dans une région donnée, on s'attend à ce qu'un petit nombre de stades phénologiques soit mentionné régulièrement, et que ces stades soient relativement fréquents. On pourra trouver des mentions d'autres stades, notamment des stades présentant une sensibilité particulière face à certains bioagresseurs, mais ceux-ci devraient, en comparaison, apparaître moins fréquemment et sans être nécessairement accompagnés des stades phénologiques voisins.

Pour cette raison, nous proposons d'introduire une métrique de confiance basée sur la fréquence du stade et de ses stades voisins.

Voisinage entre stades phénologiques La délimitation du voisinage des stades phénologiques s'apparente à un problème de recherche du plus court chemin en théorie des graphes. Les stades phénologiques sont représentés sous la forme d'un graphe dans lequel chaque stade correspond à un noeud, et reliés par des arêtes correspondant aux relations entre les stades phénologiques ($skos$:$exactMatch$, $ppdo$:$follows$...), matérialisé par le graphe de connaissances de PPDO, encodé en RDF.

Dans un premier temps, on calcule le chemin le plus court permettant de relier un noeud source à un noeud cible, c'est à dire le nombre d'arêtes à emprunter pour rejoindre le noeud cible depuis le noeud source. Dans cette version, on considérera que chaque arête empruntée augmente la distance de 1.

Ainsi, depuis le stade Eichhorn-Lorenz 07 il faut emprunter 3 arêtes pour atteindre le stade Baggiolini A. La distance est donc égale à 3 (voir figure 4).

On répète cette opération pour toutes les paires de stades phénologiques, jusqu'à obtenir une matrice des distances les plus courtes. Pour définir un voisinage il ne reste plus qu'à fixer une valeur seuil pour la distance, à partir de laquelle le stade cible n'est plus considéré comme voisin du stade source.

Derrière ce premier calcul de distance se trouve une hypothèse sous-jacente : que la distance entre deux stades est la même, quel que soit le lien qui les relie ($skos$:$exactMatch$, $skos$:$closeMatch$, $skos$:$narrowMatch$, $skos$:$broadMatch$, $ppdo$:$follows$, $ppdo$:$precedes$), la nature des stades (principal, secondaire, ou tertiaire) ou l'échelle phénologique à laquelle ils appartiennent (Baggiolini, BBCH, Eichhorn-Lorenz, IFV Labels et IFV Epicure). La sémantique du graphe n'est donc pas prise en compte.

Pour chaque bulletin on détermine ainsi un voisinage dominant, et les mentions présentant une trop grande distance par rapport au centre du voisinage dominant sont jugées peu fiables.

Hypothèse 2 : une mention de culture est plus fiable si la filière agricole à laquelle elle appartient est très présente dans le document.

Les Bulletins de Santé du Végétal peuvent être mono ou multi-cultures, mais se concentrent généralement sur l'observation de cultures appartenant à la même grande catégorie de culture (à l'exception de certains bulletins spéciaux comme les bulletins bilan ou les notes). Ainsi on trouvera des bulletins de viticulture, de maraîchage, d'arboriculture fruitière, de grandes cultures... En conséquence, si l'on repère des mentions qui n'appartiennent pas à la catégorie dominante, il sera raisonnable de penser que ces mentions sont peu fiables. Ainsi, si l'on repère une mention de culture de champignon dans un bulletin de viticulture, il y a de fortes chances que la mention soit erronée et désigne en fait un bioagresseur.

Nous proposons d'agréger les mentions de cultures qui appartiennent à la même grande catégorie de culture (fcu:$Grandes\_categories$) afin de mesurer la fréquence conceptuelle de ces catégories à l'intérieur de chaque BSV. Chaque bulletin est donc associé à une liste des catégories de cultures ainsi qu'aux fréquences conceptuelles de ces catégories à l'intérieur du bulletin. Nous fixons ensuite un seuil minimum pour la fréquence conceptuelle en-dessous duquel les mentions rattachées à la catégorie sont enlevées. Ainsi dans la phrase *«Gel dans la nuit du 12/04 au*

9. La forme gel peut correspondre à une culture fcu:$Gels$, définie comme *«Ensemble de plantes apparaissant sur un terrain agricole non mis en culture (jachère).»*

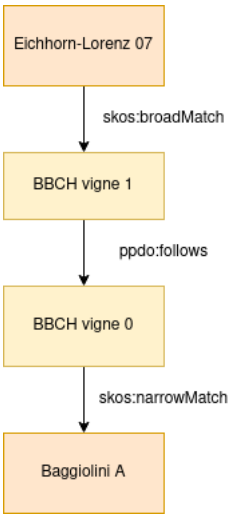

FIGURE 4 – Distance entre les stades Eichhorn-Lorenz 07 et Baggiolini A.

13/04» "gel" est initialement identifié comme une culture et liée à fcu:$Gels$. La désambiguïsation permet d'identifier que la catégorie fcu:$Plantes\_non\_recoltees$ à laquelle fcu:$Gels$ appartient est trop faiblement représentée dans le bulletin et d'enlever la mention erronée.

Lorsqu'une mention appartient à plusieurs catégories (pour une plante multi-usages), celle-ci est conservée tant qu'au moins l'une de ces catégories est fortement représentée dans le bulletin. Ainsi dans l'un des documents on trouve une mention "ananas". Dans ce document précis, la mention est conservée même si elle appartient à la catégorie fcu:$Cultures\_tropicales$ car elle est également rattachée à fcu:$Cultures\_fruitieres$ qui elle est suffisamment présente dans le document pour qu'on conserve ses mentions.

Puisque les bulletins décrivent des observations sur des cultures appartenant à la même catégorie, on pourrait se demander si il ne serait pas plus judicieux de ne conserver que les mentions de la catégorie dominante. Deux raisons nous poussent à ne pas adopter cette méthode :

— Certaines mentions ambiguës comme "champignons" sont parfois très fréquentes et peuvent biaiser la catégorie dominante, au détriment du type de culture réel du bulletin.

— Certaines mentions ne décrivent certes pas des cultures qui ont fait l'objet d'observations, mais représentent bien des mentions de cultures, par exemple on trouvera dans un bulletin viticole, dans un paragraphe qui décrit l'action d'un bioagresseur *«Les fruits les plus attaqués sont l'avocat, la mangue et la papaye mais l'espèce s'en prend aussi au citron, goyave, banane, nèfle du Japon, tomate, cerise de Cayenne, fruit de la passion, kaki, ananas, pêche, poire, abricot, figue et café»*. Si possible, on souhaiterait une méthode qui préserve ces mentions.

3.4 Résultats

La table 1 présente les résultats pour les stades phénologiques. On peut voir qu'un très grand nombre de mentions

	Mentions		Stades phénologiques	
	Corpus	Document	Corpus	Document
Annotation	4375	57	129	21,4
Désambiguïsé	4139	53,8	127	20,5

TABLE 1 – Annotation des stades phénologiques : nombres de mentions et stades phénologiques distincts annotés (dans le corpus et par bulletin), avant et après désambiguïsation.

	Mentions		Cultures	
	Corpus	Document	Corpus	Document
Annotation	1789	23.2	61	5.5
Désambiguïsé	1640	21.3	49	4.4

TABLE 2 – Annotation des cultures : nombres de mentions et cultures distinctes annotées (dans le corpus et par bulletin), avant et après désambiguïsation.

est relevé, ceci est dû en partie aux alignements entre les différentes échelles, qui peuvent multiplier une annotation par cinq si un équivalent existe dans chaque échelle. Les annotations proviennent en majorité des alignements (59,3%) puis de la projection des étiquettes (29,3%) et enfin des motifs (11,4%). Les annotations trop éloignées du voisinage dominant représentent 5.4% des annotations de départ, mais font peu baisser le nombre de stades phénologiques distincts repérés dans le corpus.

Le tableau 2 présente les résultats de l'annotation sur les cultures. Un nombre relativement importants de cultures est détecté dans le corpus, avec 1789 mentions et 61 cultures distinctes. Ce chiffre paraît important étant donné l'homogénéité du corpus, qui ne contient que des bulletins sur la vigne. Après désambiguïsation le nombre de cultures distinctes baisse de façon importante (20%) comparé au nombre de mentions supprimées (10%), ce qui signifie que les mentions appartenaient probablement à des concepts peu fréquents, ce qui semble encourageant au vu des exemples d'ambiguïtés présentés dans la section 3.2

qu'on imagine pour la plupart plutôt ponctuels.

4 Travaux connexes

Annotation sémantique d'entités pour l'épidémiosurveillance D'autres projets proposent une annotation sémantique pour l'épidémiosurveillance en santé végétale. C'est le cas du projet TIERS-ESV [10] qui met en place une chaîne de traitement pour faciliter l'automatisation de la veille sanitaire. Au sein de ce projet la reconnaissance et normalisation des entités nommées se concentre sur les mentions d'organismes nuisibles, leurs végétaux hôtes et leurs vecteurs, ainsi que la date et le lieu de l'observation. Les méthodes adoptées sont similaires, à savoir une extraction d'information basées sur des ressources structurées.

Métriques de désambiguïsation basées sur les relations entre concepts En recherche d'information, la désambiguïsation de concept permet de sélectionner dans une ressource sémantique (ontologie ou modèle SKOS) le bon concept à relier à un terme extrait par rapport à un ensemble de candidats. Les concepts ainsi sélectionnés peuvent être utilisés pour caractériser le contenu sémantique du document. Les scores de confiances proposés dans la section 3.3 partagent des similarités avec les métriques de désambiguïsation présentés dans [2]. L'idée générale consiste à exploiter les relations entre concepts pour désambiguïser les mentions. Ainsi une mesure comme la fréquence conceptuelle prendra en compte la fréquence des mentions des concepts fils. Notre approche est un peu différente car nous utilisons ces scores de confiance non pas pour relier la mention au bon concept, mais pour décider si la mention appartient bien au type d'entité d'intérêt (culture ou stade phénologique).

5 Conclusion et perspectives

Nous avons présenté notre chaîne de traitement pour l'annotation sémantique d'un corpus de la vigne des Bulletins de Santé du Végétal. La reconnaissance et la normalisation des entités nommées de type $Culture$ et $Stade\ Phnologique$ dans les Bulletins de Santé du Végétal repose en grande partie sur deux ressources sémantiques : le thésaurus French Crop Usage (FCU) et le graphe de connaissances peuplant BBCH-based Plant Phenological Description Ontology (PPDO). Nous avons également expérimenté deux méthodes pour améliorer la qualité des annotations, au travers d'une désambiguïsation qui repose sur la connectivité entre les entités dans les ressources sémantiques notamment sur leur voisinage (pour les stades phénologiques) et sur leur ancêtre commun (pour les cultures).

À l'heure actuelle, une évaluation systématique de ces annotations ainsi que des méthodes de désambiguïsation reste encore à fournir. Cet effort est en cours puisqu'un corpus de référence des BSV viticoles, annoté sur les cultures, stades phénologiques et bioagresseurs est en cours de développement. Ce corpus de référence pourra servir de jeu de données de test pour évaluer les méthodes proposées et être réutilisé par la communauté.

10. https ://plateforme-esv.fr/node/24638

Remerciements

Ce travail a été réalisé avec le soutien du projet "Des Données aux Connaissances en Agronomie et Biodiversité (D2KAB–www.d2kab.org) financé par l'Agence Nationale de la Recherche (ANR-18-CE23-0017) . Nous remercions également les membres de la tâche 4.3 du projet D2KAB : Sophie Aubin, Sonia Bravo, Anna Chepaikina, Baptiste Darnala, Catherine Faron, Matthieu Hirschy, Clement Jonquet, Franck Michel, et Nadia Yacoubi. Un remerciement particulier pour les ingénieurs spécialistes de la vigne qui nous ont aidé dans ce travail : Arnaud Charleroy de INRAE laboratoire MISTEA, Thierry Lacombe de SupAgro et Xavier Delpuech de l'Institut Français de la Vigne et du Vin (IFV).

Références

[1] Mouhamadou Ba and Robert Bossy. Interoperability of corpus processing work-flow engines : the case of alvisnlp/ml in openminted. In *Proceedings of the Workshop on Cross-Platform Text Mining and Natural Language Processing Interoperability (INTEROP 2016) organised with LREC 2016*, pages pp. 15–18, Portorož, Slovenia, 2016.

[2] Mustapha Baziz, Mohand Boughanem, Nathalie Aussenac-Gilles, and Claude Chrisment. Semantic cores for representing documents in ir. In *Proceedings of the 2005 ACM symposium on Applied computing*, pages 1011–1017, 2005.

[3] Uwe Meier. *Growth stages of mono- and dicotyledonous plants : BBCH Monograph*. Open Agrar Repositorium, 2018.

[4] Alistair Miles and Sean Bechhofer. *SKOS Simple Knowledge Organization System Reference*. W3C Recommendation. World Wide Web Consortium, United States, August 2009.

[5] Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. Stanza : A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics : System Demonstrations*, 2020.

[6] Catherine Roussey, Xavier Delpuech, Florence Amardeilh, Stephan Bernard, and Clement Jonquet. Semantic description of plant phenological development stages, starting with grapevine. In Emmanouel Garoufallou and María-Antonia Ovalle-Perandones, editors, *Metadata and Semantic Research*, pages 257–268, Cham, 2021. Springer International Publishing.