Annotation des Bulletins de Santé du Végétal

Anna Chepaikina
Robert Bossy
Catherine Roussey
Stéphan Bernard

01.03.2021 - à présent

Introduction

« Construction et enrichissement des jeux de données FAIR à travers des techniques de text-mining »

Projet D2KAB

Data to Knowledge in Agronomy and Biodiversity

2019-2023

www.d2kab.org

n° 2
16 avril 2019

Viticulture

Bulletin de santé du végétal ÉCOPHYTO

À retenir cette semaine

Épisode de gel les 13, 14 et 15 avril. Dégâts à estimer dans les jours à venir

Le débourrement se poursuit mais est au ralenti. Stade « pousse verte » majoritaire

Quelques dégâts de mange-bourgeons, pour l'instant sans conséquence

Réseau des Chambres d'Agri-

Crédit photo :

BSV réalisé en fonction des observations de la situation sanitaire des vignobles à partir des données des vignobles suivis dans le cadre du réseau de parcelles en Auvergne-Rhône-Alpes. Observations effectuées par les membres du réseau BSV en application du protocole harmonisé national d'observations. Cette année, le réseau comprend 22 parcelles observées par 14 observateurs sur 5 cépages différents.

Données du réseau

15 parcelles renseignées, 12 dans le vignoble de Saint Pourçain, 3 dans le vignoble des Côtes d'Auvergne.

Stades phénologiques

t°c minimale	Besson (03)	Saulcet (03)	Chateaugay (63)	Boudes (63)
05/04/2019	-2,1	-1,9	-2,6	-4,9
13/04/2019	-1	-1,1	-1,1	-2,5
14/04/2019	-3,1	-3,7	-2,6	-3,8
15/04/2019	-3,8	-3,5	-3,3	-4,7

Pousse

La température est encore descendue en dessous de 0 cette fin de semaine. Même si l'on voit des bourgeons touchés sur quelques parcelles plus gélives, les dégâts sont en moyenne relativement peu importants au vu des relevés de températures ! Les jours prochains permettront de mieux estimer l'impact réel de cet épisode de gel (superficiel ou gel ayant atteint les embryons d'organes)

.

Photo C.Peignelin

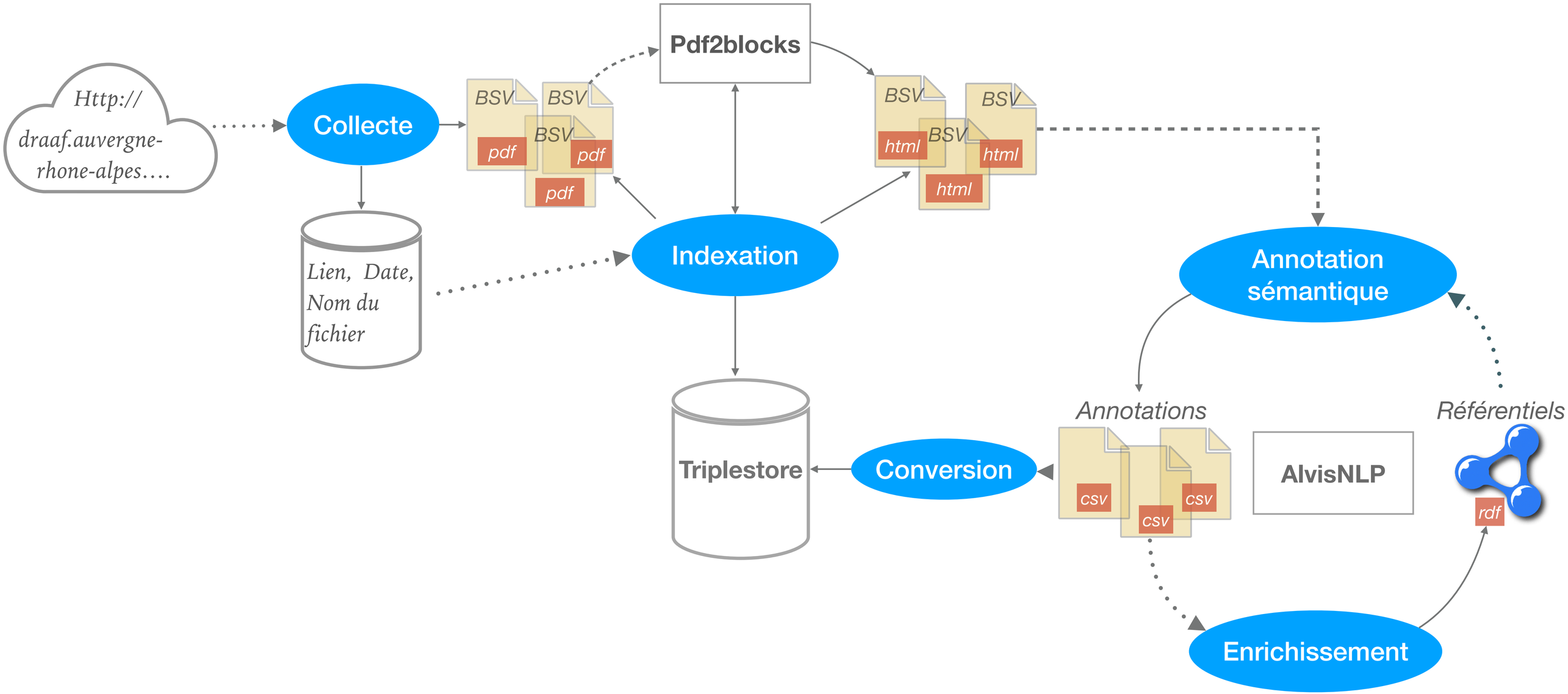
AGRICULTURES & TERRITOIRES
CHAMBRES D'AGRICULTURE AUVERGNE-RHÔNE-ALPES

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE
MINISTÈRE DE L'AGRICULTURE ET DE L'ALIMENTATION

Liberté • Égalité • Fraternité
RÉPUBLIQUE FRANÇAISE
MINISTÈRE DE LA TRANSITION ÉCOLOGIQUE ET SOLIDAIRE

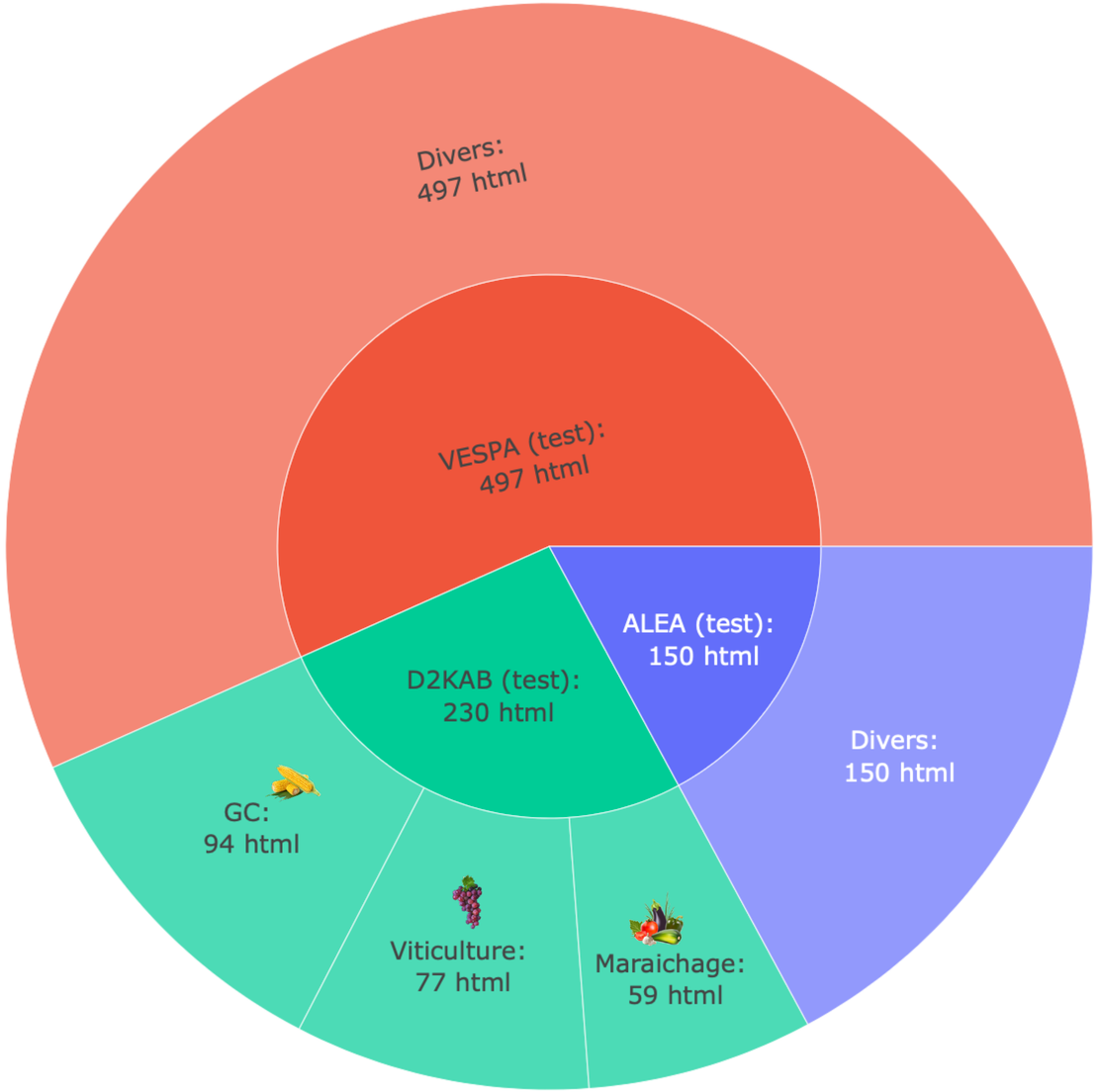
Exemple d'un Bulletin de Santé du Végétal au format pdf

Schéma global



Corpus

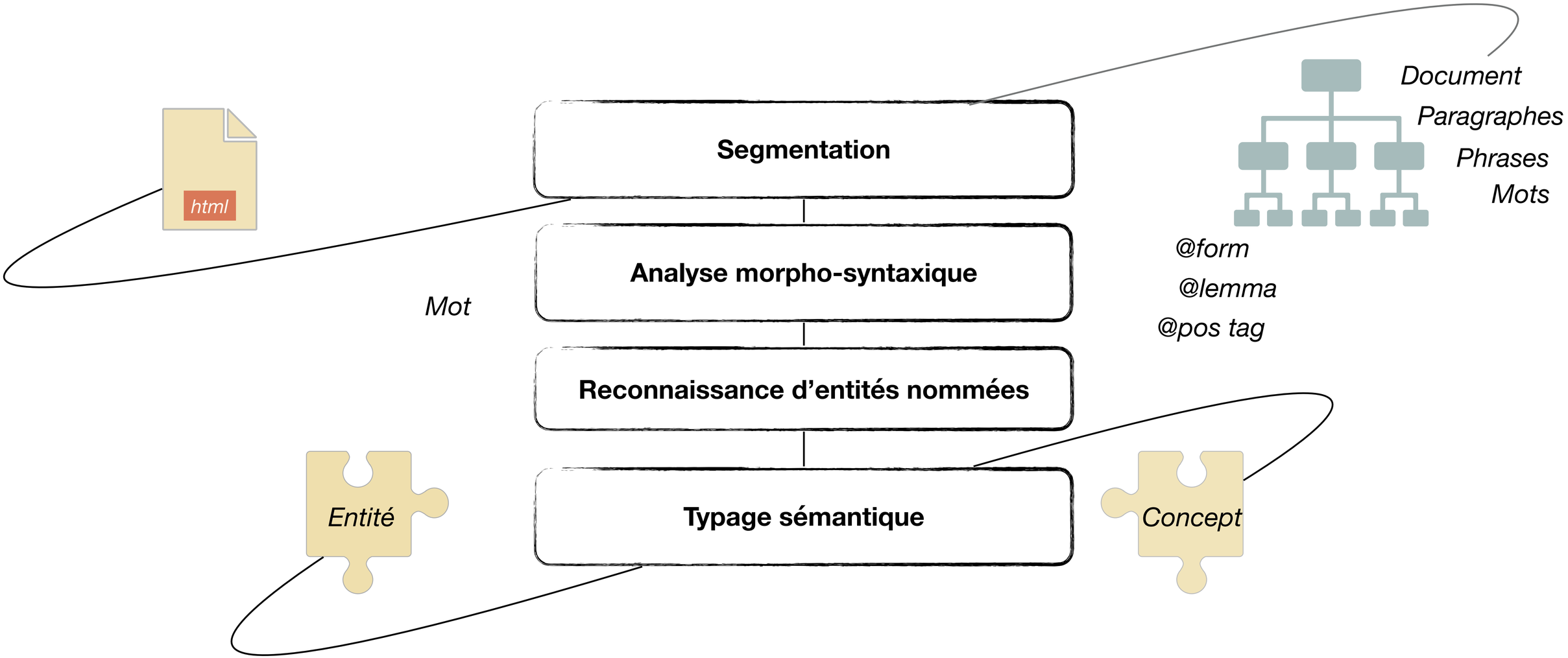
- **Corpus de test D2KAB** (230 BSV)
  - 3 cultures (Viticulture - Grandes Cultures - Maraîchage)
  - 17 regions (après la réforme)
  - 2019
- **Corpus de test VESPA** (497 BSV)
  - une demi-douzaine d'éditions (Arboriculture - Grandes Cultures …) par région
  - 27 régions (avant la réforme)
  - 2009-2016
- **Corpus de test ALEA** (150 BSV)
  - tirés aléatoirement parmi les BSV collectés automatiquement



Répartition des bulletins par culture

Outils

AlvisNLP est un moteur de workflow d'annotation automatique de corpus, développé par l'équipe de Bibliome et intégrant différents outils de traitement du langage naturel :

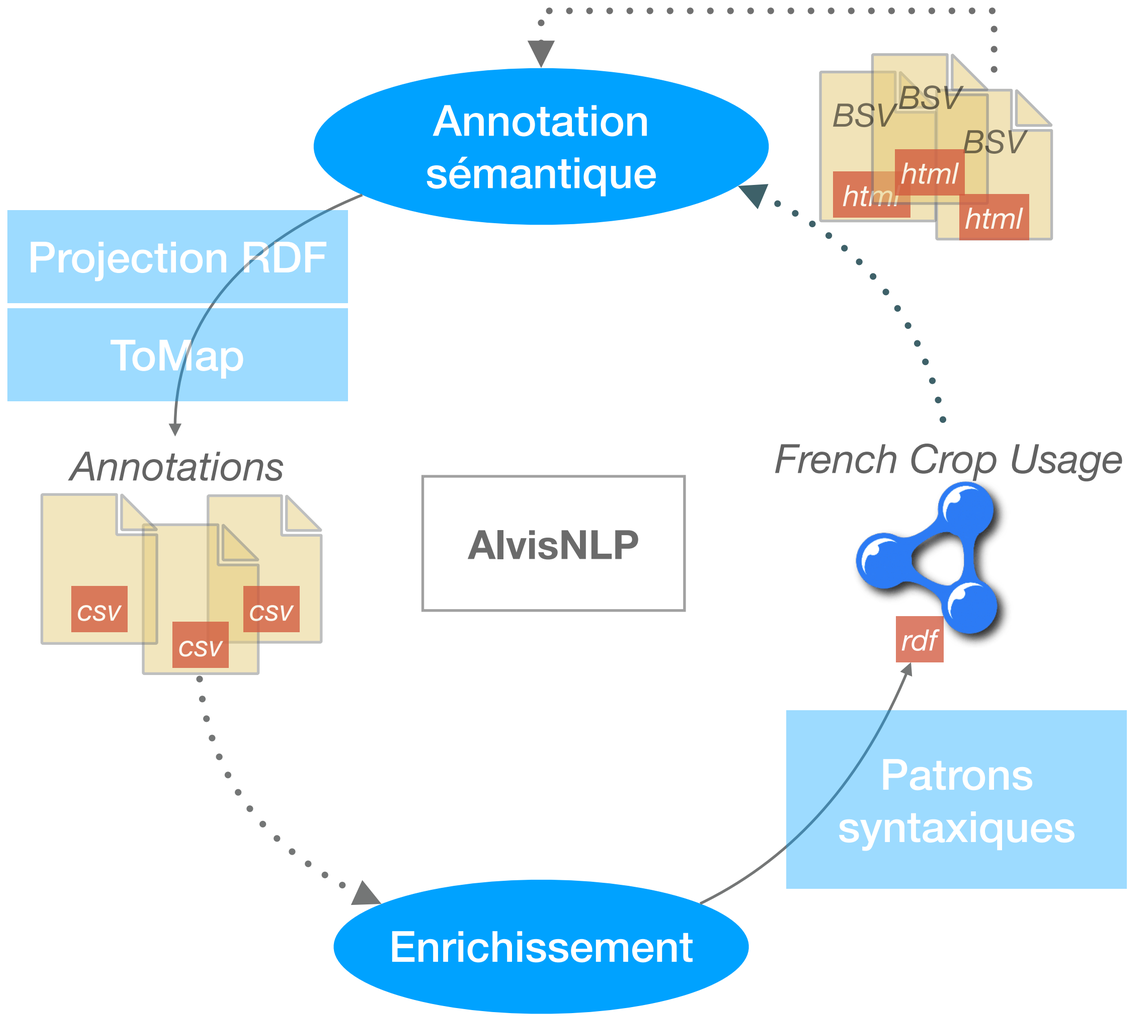


Workflow

Référentiel « French Crop Usage »

http://ontology.inrae.fr/frenchcropusage

- multi usages
- usages des plantes cultivées
  - grandes catégories de cultures
    - Arboriculture
    - Cultures aromatiques medicinales parfum
    - Cultures fourrageres
    - Cultures fruitieres
    - Cultures legumieres champignons
    - Cultures tropicales
    - Grandes cultures
    - Horticulture ornementale
    - Plantes non recoltees
    - Semences
    - Zones non agricoles



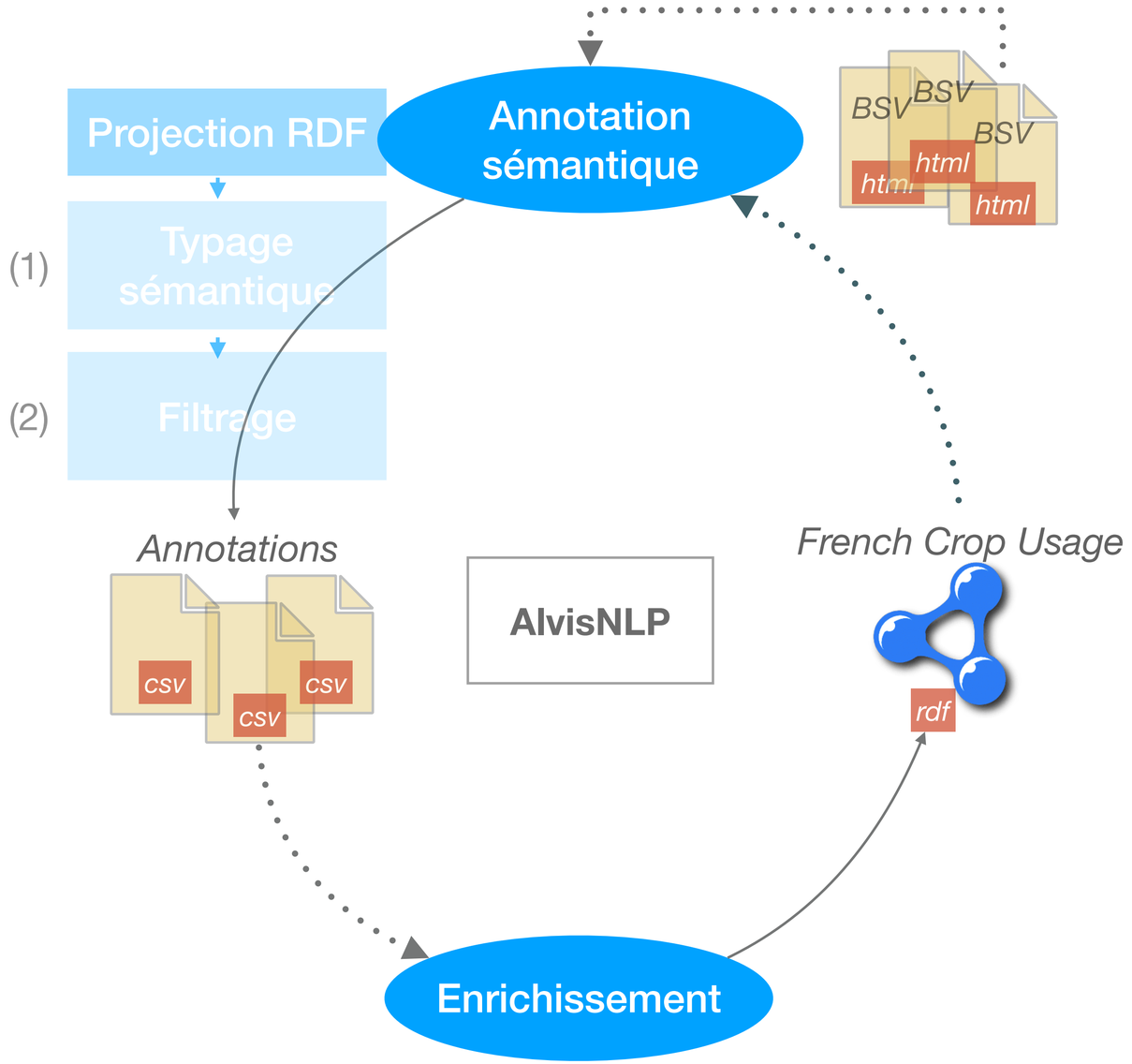
Annotation

Projection RDF

<u>Projection RDF</u> est une méthode d'association des entités nommées avec des étiquettes de concepts venant de terminologies skos et ontologies owl

```
<project-baseline class="RDFProjector">

  <!-- référentiel -->
  <source>resources/fcu/frenchCropUsage_20210525.rdf</source>
  <language>fr</language>
  <uriFeatureName>uri</uriFeatureName>
  <resourceTypeURIs>owl:NamedIndividual</resourceTypeURIs>

  <!-- annotations sur le corpus des bsv -->
  <subject layer="words" feature="lemma"/>
  <targetLayerName>fcu-baseline</targetLayerName>
  <constantAnnotationFeatures>type=RDFProjector
  </constantAnnotationFeatures>

  <!-- traitements -->
  <allowJoined>true</allowJoined>
  <joinDash>true</joinDash>
  <caseInsensitive>true</caseInsensitive>
  <ignoreDiacritics>true</ignoreDiacritics>

</project-baseline>
```

Extrait d'un plan AlvisNLP

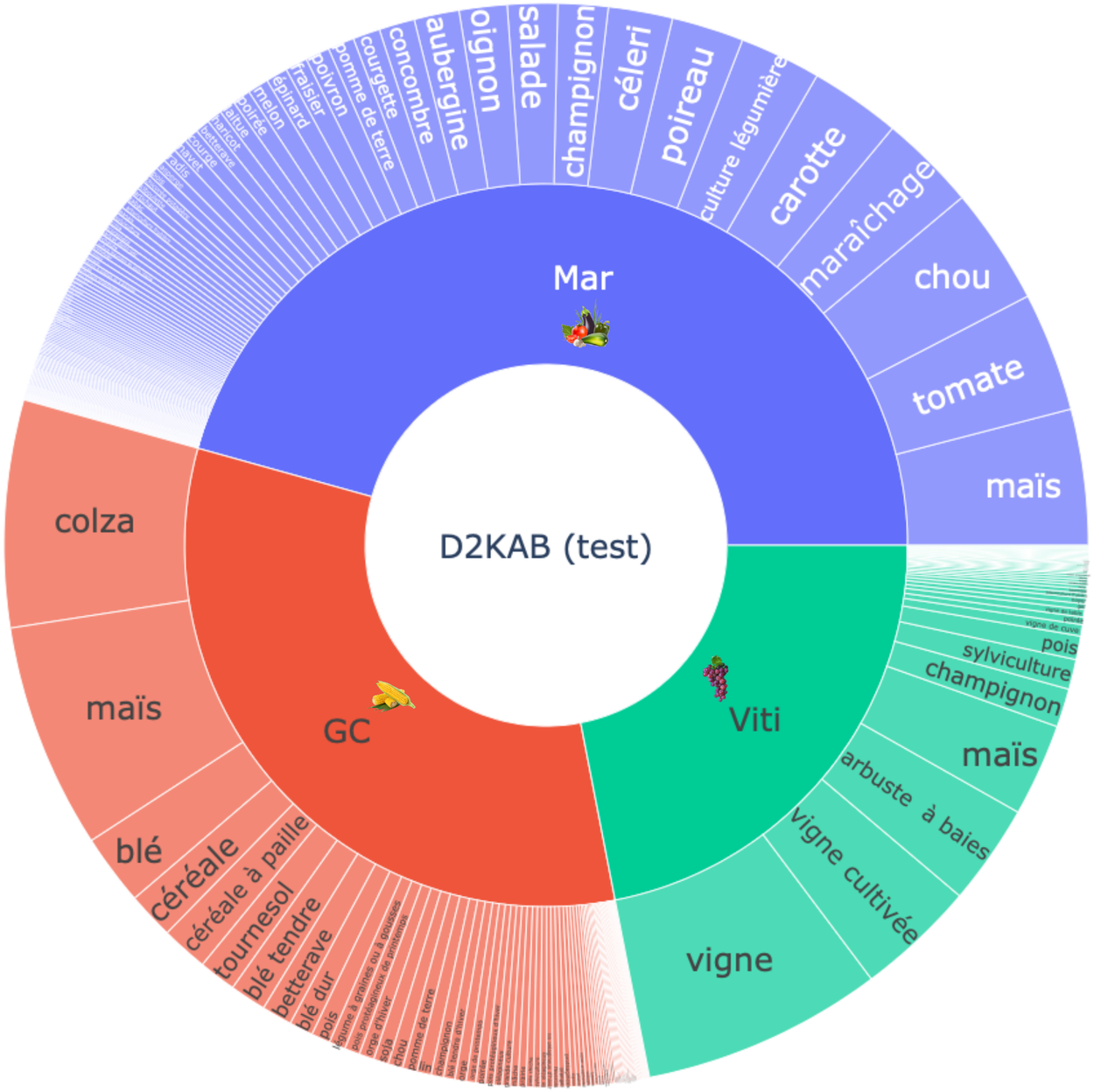


Corpus	Nb de mots/ doc	Nb de doc/ corpus	Nb de mentions/ corpus	Nb de concepts/ corpus	Couverture du corpus	Densité d'entités
D2KAB (Test)	2850	230	9397	189	100 %	1.43 %
VESPA (Test)	1815	497	12779	224	89.54 %	1.59 %
ALEA (Test)	2981	150	6662	195	100 %	1.48 %

Résultats de la projection par corpus

Concepts filtrés :

- fruit
- fleur
- semence
- côte
- soleil
- gel
- orange
- marron



Répartition des concepts trouvés dans le corpus de test D2KAB

Annotation

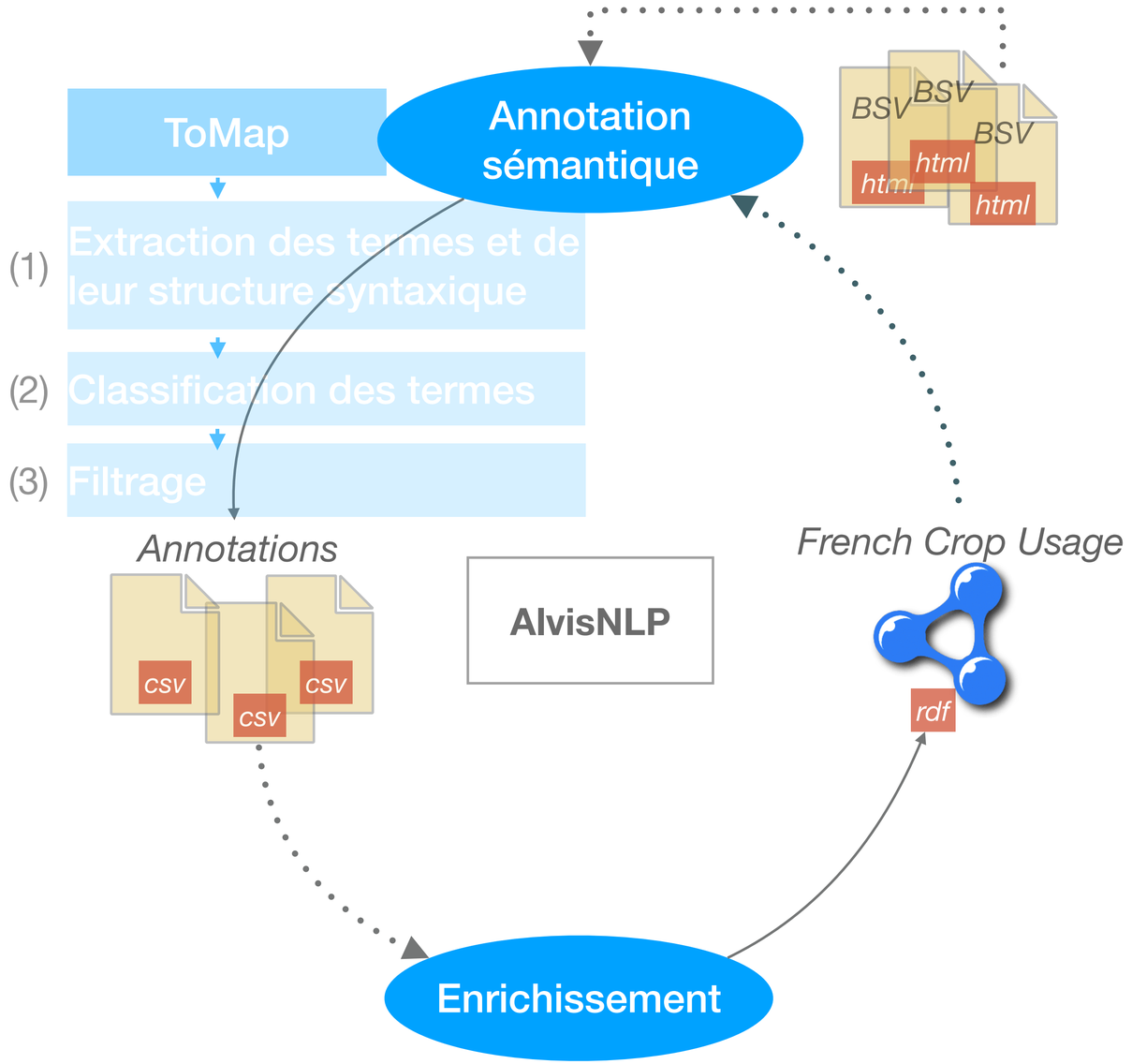
ToMap

ToMap est une méthode non-supervisée de classification des termes en comparant la structure syntaxique du terme et les étiquettes des concepts (catégories)

Avantages : typage sémantique des entités nommées de structure syntaxique variable.

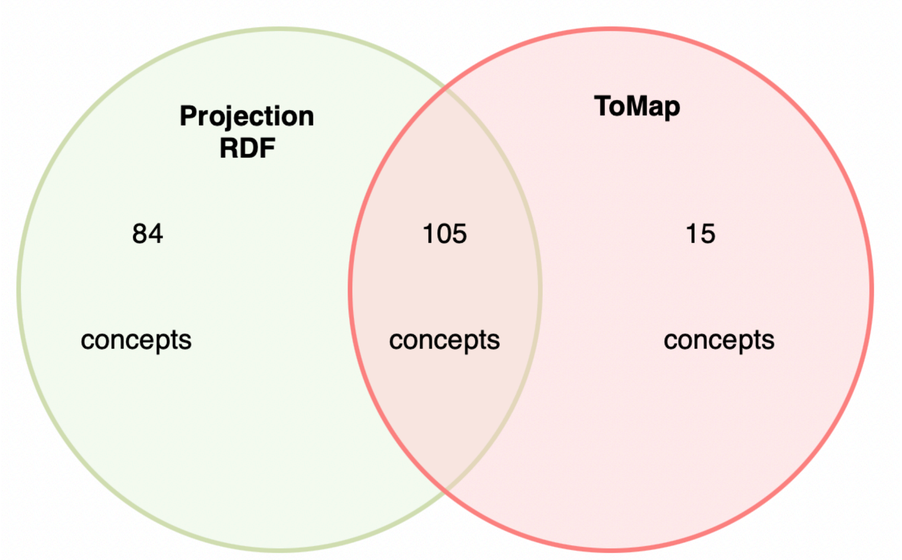
Exemple :
« Feuilles fermées »
- - - > « Fermeture des feuilles »

Publications : « Event extraction of bacteria biotopes: a knowledge-intensive NLP-based approach »

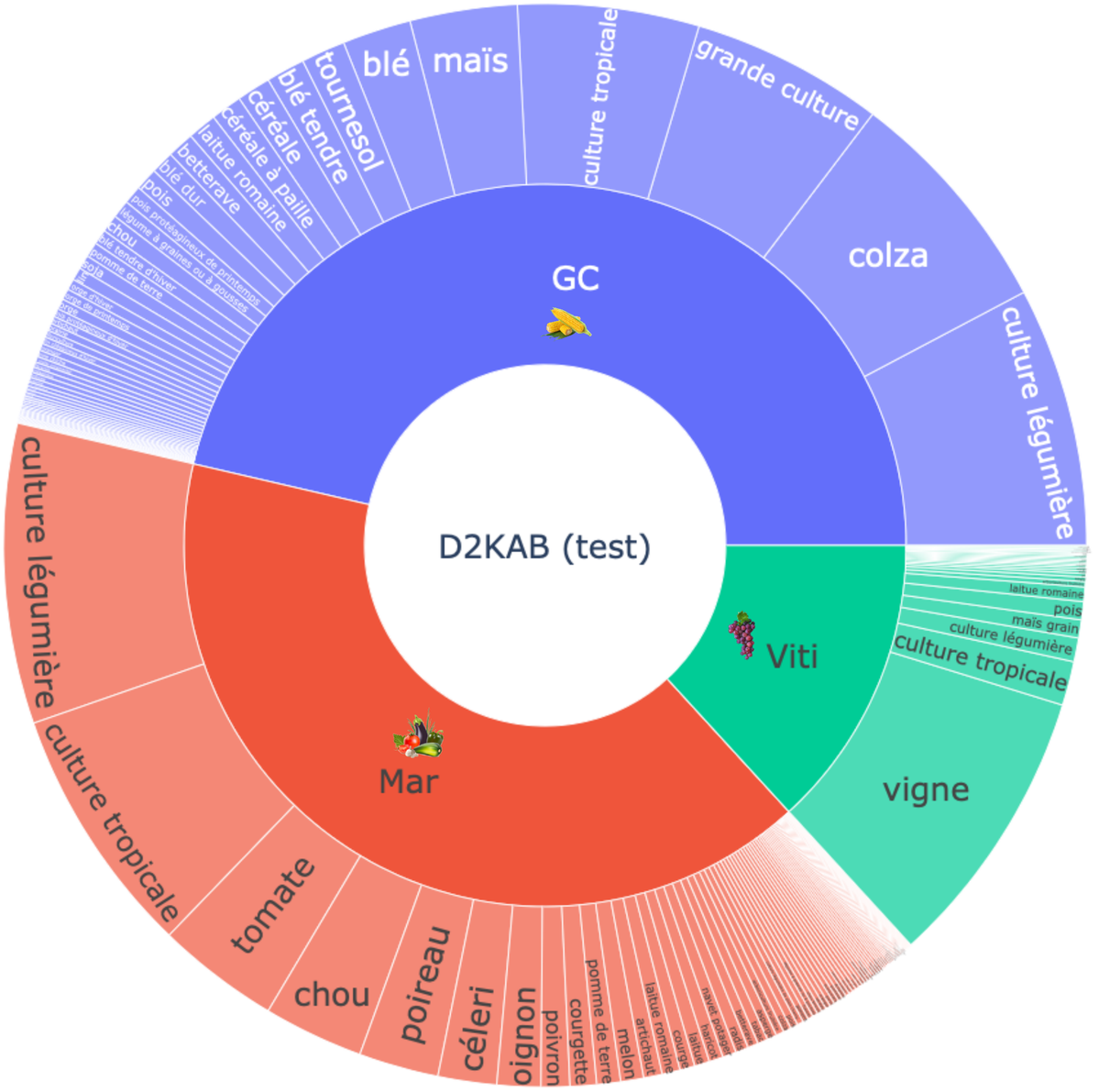


Corpus	Nb de mots/ doc	Nb de doc/ corpus	Nb de mentions/ corpus	Nb de concepts/ corpus	Couverture du corpus	Densité d'entités
D2KAB (Test)	2850	230	7599	120	100 %	1.15 %
VESPA (Test)	1815	497	9942	148	89.13 %	1.23 %
ALEA (Test)	2981	150	5220	128	100 %	1.16 %

Résultats de la projection par corpus



Concepts trouvés par la projection RDF vs. ToMap



Répartition des concepts trouvés dans le corpus de test D2KAB

Annotation

Keywords

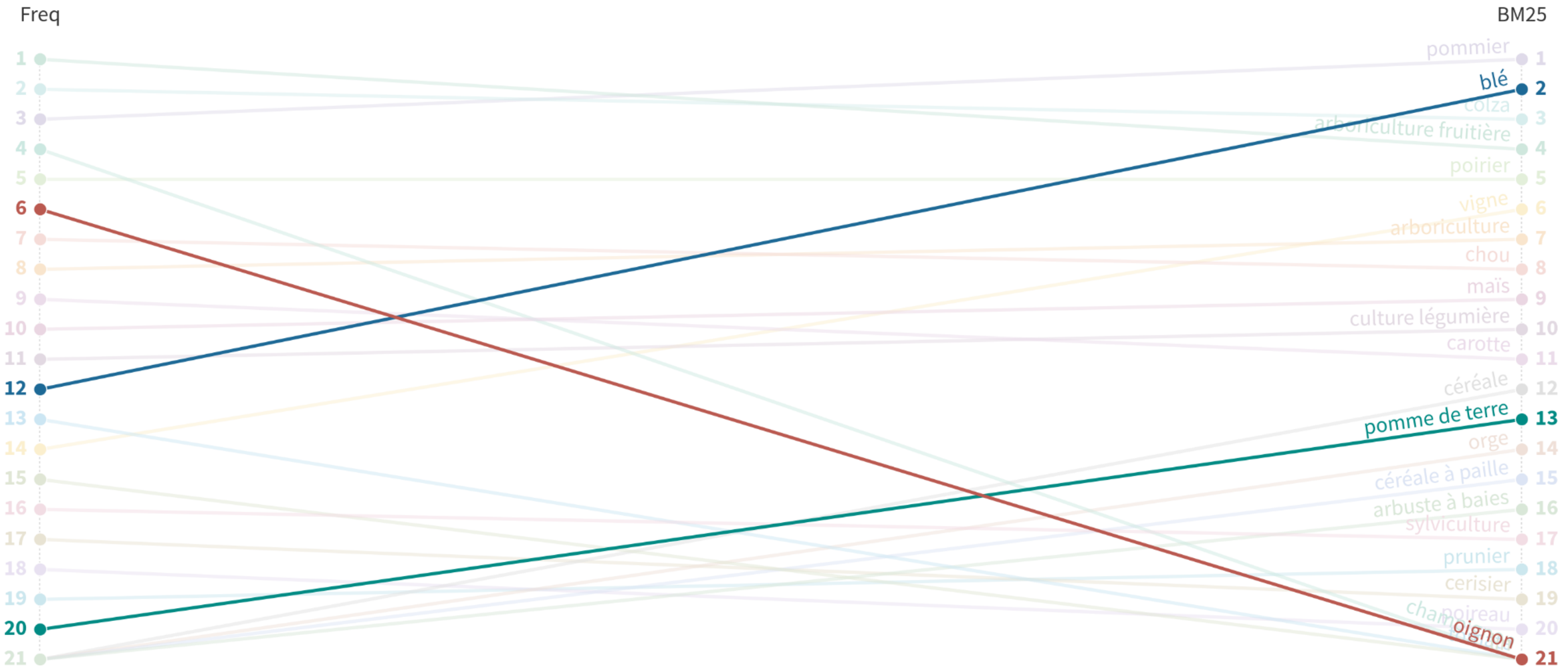
Objectif : sélectionner les concepts les plus saillants pour l'indexation.

- ~~Nombre d'occurrences~~
- ~~TFIDF~~ [1]
- **Okapi BM25** [2]
    - Modèle probabiliste
    - Prise en compte de la longueur du document

---

1. *https://doi.org/10.1007/978-0-387-30164-8_832*
2. *https://doi.org/10.1007/978-0-387-39940-9_921*

Term frequency vs. Okapi BM25



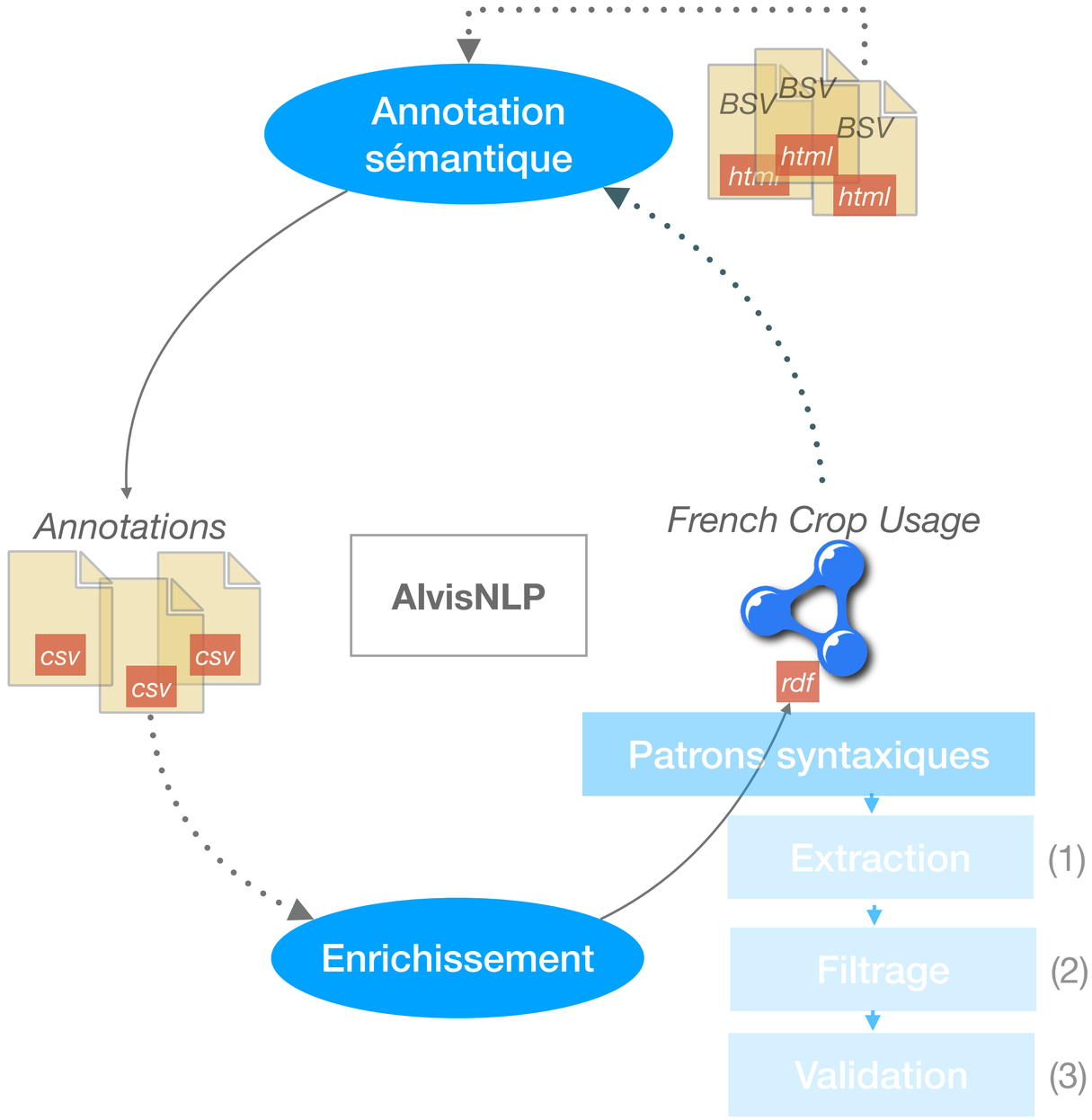
Enrichissement

Patrons syntaxiques

Extraction des patrons syntaxiques dans AlvisNLP demande à appliquer une expression régulière sur une séquence d'annotations

```
<N_PUN_N_KON_N class="PatternMatcher">
  <pattern>
    (
      [overlapping:GN]
      [ @form == ","]
    )+
    [overlapping:GN]
    [ @form == "et"]
    [overlapping:GN]
  </pattern>
  <actions>
    <createAnnotation layer="conjunctions"/>
  </actions>
  <constantAnnotationFeatures>
    type=N_PUN_N_KON_N
  </constantAnnotationFeatures>
</N_PUN_N_KON_N>
```

Extrait d'un plan AlvisNLP



bioagresseurs secondaires comme les pourritures et les drosophiles. Les fruits sont alors non commercialisables. Extrêmement polyphage elle s'attaque à plus de 300 plantes hôtes, plantes cultivées et sauvages, légumières ou fruitières. 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. Les légumes concernés sont notamment les tomates, poivrons, melons et courges. Comme les autres mouches de cette famille, elle a un cycle de vie très court et une fécondité élevée. La femelle peut pondre entre 800 à 1 500 œufs durant sa vie à raison d'une vingtaine par jour.

Chaume
Cerise de Cayenne
Nèfle du Japon
Épinard de printemps
Épinard d’automne
Oignons de semis
Échalotes de plantation
Aillet
Romanesco
Saule, hêtre, sycomore, frêne, bouleau des arbres …

Problématiques

- Segmentation
- Lemmatisation
- Concepts ambigus

Française vs. Anglaise

@form = Fraise

@lemma = Fraiser?

« Champignon » (Culture ou Bioagresseur ?)

« Gel » (jachère ou gel ?)

« Côte » (Côte d'Azure ou une partie des plantes ?)

Fruits ou couleurs ?

Améliorations

- Segmentation sémantique

  - Meta, édition, footers, sommaire, réseau d’observation, pestes, stades phéno …

```
<liste_ravageurs class="PatternMatcher">
    <layerName>html</layerName>
    <pattern>
        [ @tag ^= "H" and @tag =~ "[0-9]" and  str:lower(@form) =~ "ravageur"]
        [ @tag ^= "H" and @tag =~ "[0-9]" and overlapping:bioagressors or
        @tag == "P" and overlapping:bioagressors]{1,5}
    </pattern>
    <actions>
        <createAnnotation layer="themes"/>
    </actions>
    <constantAnnotationFeatures>type=RISQUE</constantAnnotationFeatures>
    <overlappingBehaviour>ignore</overlappingBehaviour>
</liste_ravageurs>
```

Extrait d’un plan AlvisNLP

- Context temporel et spatial
- Hiérarchie parent-child (conjonctions)
- Paramétrage de ToMap

Merci pour votre attention !