Comment la fouille de texte appliquée aux listes d'ingrédients contribue à analyser l'offre alimentaire ?

Tristan Salord, Marie-Benoît Magrini

Comment la fouille de texte appliquée aux listes d'ingrédients contribue à analyser l'offre alimentaire ?

Premières investigations sur les innovation-produits contenant des légumineuses de la base MINTEL-GNPD

Tristan SALORD, Marie-Benoît MAGRINI

UMR AGIR, INRAE Occitanie – Toulouse
Équipe ODYCEE

Colloque Végétransform, Nantes, 3 juin 2022

Projet Région KING
(2021-2023)

Constats

- Peu/pas d'études scientifiques approfondie des innovations produits alimentaires sur les marchés au regard de leur composition *« there is a lack of information in the scientific literature on type of ingredients used in packaged foods » Ahuja et al., 2021*
- … et croisant leur positionnement de marché en termes de catégories produits, circuits de distribution et mentions valorisantes associées

→ Encore moins sur les produits utilisant des légumineuses !

- Défaut d'algorithmes de décomposition des listes d'ingrédients, de dictionnaires partagés sur la taxonomie des espèces et variétés, sur la taxonomie des profils d'ingrédients au regard des procédés techniques subis

→ Etablir ces données est essentiel pour ces analyses et construire des trajectoires d'innovation dans l'agroalimentaire, identifier les modèles agroalimentaires associés, évaluer aussi la biodiversité (marchande)…

Cusworth et al. (2021 "Legume dreams" :7) highlights two main pathways of development of legumes :
"They [legumes] can be the staples of the eco-pessimist who would prefer the revitalisation of unprocessed, traditional and whole foods to help minimise the negative environmental externalities of food production. They also serve as the fungible base materials for high-tech food processors who aim to produce convincing meat-simulacra to displace unsustainable meat and dairy products…"

Constats

- Impossibilité actuelle de coupler MINTEL/KANTAR pour identifier les produits qui survivent sur le marché, et donc comment cette offre trouve une demande

→Quels accès à venir du code-barre dans les données KANTAR ?

→ *D*évelopper des techniques de reconnaissance textuelle à partir du nom des produits co-repérés dans MINTEL/KANTAR…

- **Au préalable, il reste nécessaire d'identifier le profil des ingrédients et leur formulation**

FOODCOP : un nouveau parser des listes d'ingrédients

Logiciel déposé auprès de l'APP (DI-RV-21-0105) sous licence Creative Commons V4, Salord et al., 2021

disponible en ligne sous condition de citation (CC4):
https://github.com/Pythrix/FOODCOP

Data in Brief 42 (2022) 108173

Contents lists available at ScienceDirect

Data in Brief

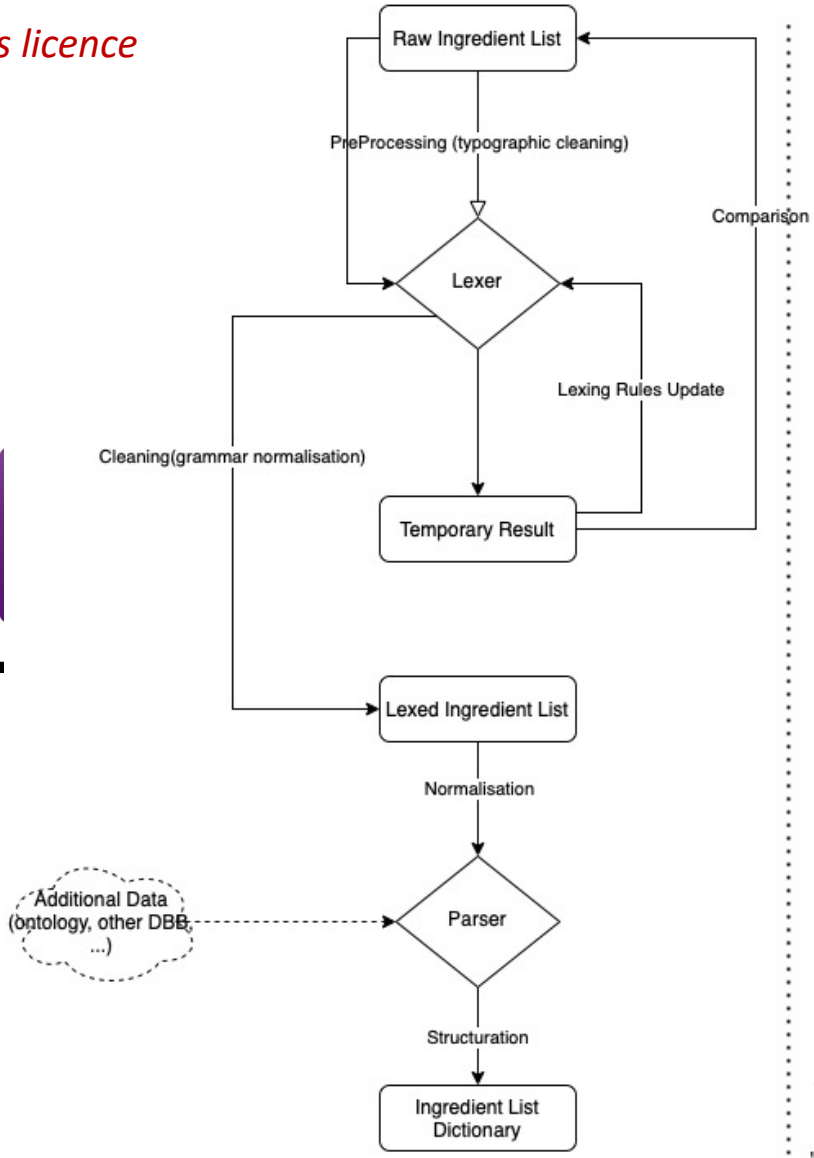
journal homepage: www.elsevier.com/locate/dib

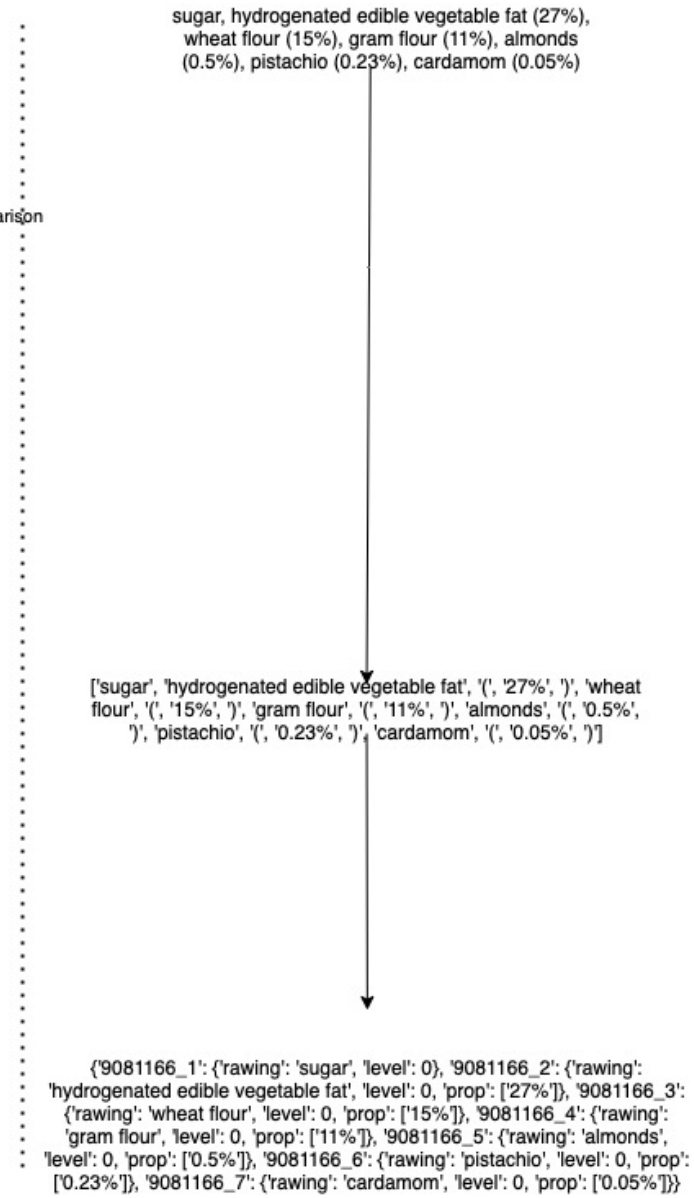
Data Article

Packaged foods with pulse ingredients in Europe: A dataset of text-mined product formulations

Tristan Salord [a,*], Marie-Benoît Magrini [a], Guillaume Cabanac [b]

[a] AGIR, INRAE, University Toulouse, Castanet-Tolosan, France
[b] CNRS, IRIT, University Toulouse, Toulouse, France





Un parser organisant la liste en dictionnaires structurés

```
'''Water,milk protein concentrate, vegetable oil (canola oil, high oleic
sunflower oil, corn oil), soy protein isolate (2%), calcium caseinate(2%),
sodium(2%) caseinate(2%), vitamins and minerals (potassium citrate,
magnesium phosphate, vitamin B12)'''
```

Distinction des rangs d'apparition (Rank)

Distinction des profondeurs d'apparition (Depth)

. Ingredient lists of multiple level.

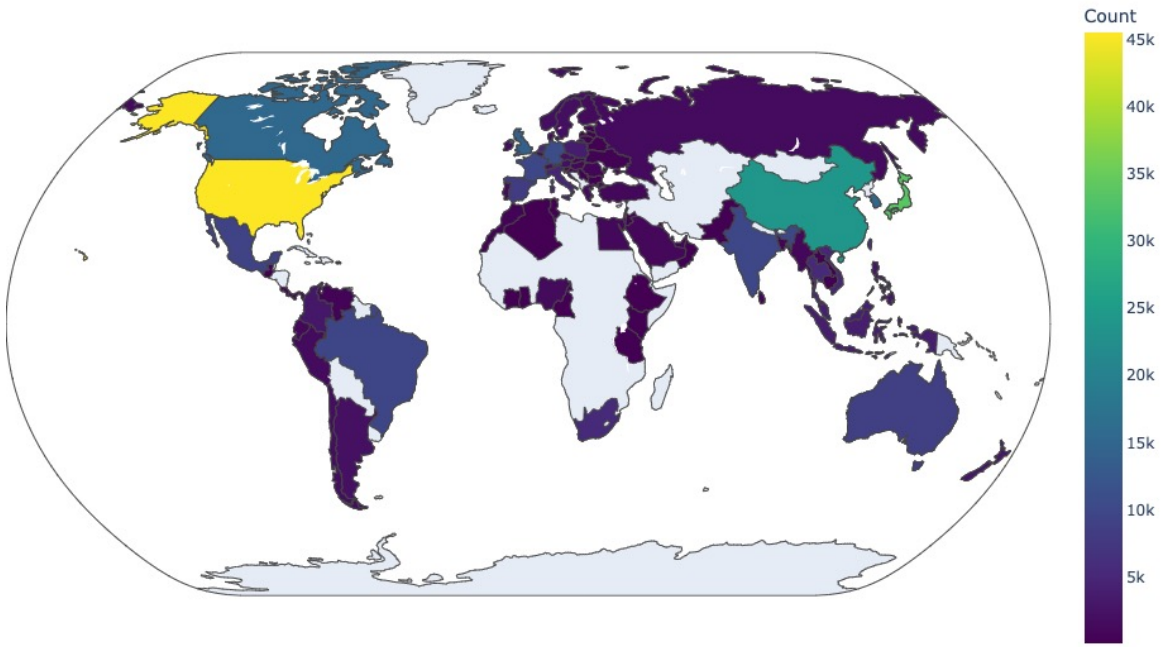
```
{'15_1': {'rawing': "'water", 'level': 0},
 '15_2': {'rawing': 'milk protein concentrate', 'level': 0},
 '15_3': {'rawing': 'vegetable oil', 'level': 0},
 '15_4': {'rawing': 'canola oil', 'level': 1},
 '15_5': {'rawing': 'high oleic sunflower oil', 'level': 1},
 '15_6': {'rawing': 'corn oil', 'level': 1},
 '15_7': {'rawing': 'soy protein isolate', 'level': 0, 'prop': ['2%']},
 '15_8': {'rawing': 'calcium caseinate', 'level': 0, 'prop': ['2%']},
 '15_9': {'rawing': 'sodium', 'level': 0, 'prop': ['2%']},
 '15_10': {'rawing': 'caseinate', 'level': 0, 'prop': ['2%']},
 '15_11': {'rawing': 'vitamins and minerals', 'level': 0},
 '15_12': {'rawing': 'potassium citrate', 'level': 1},
 '15_13': {'rawing': 'magnesium phosphate', 'level': 1},
 '15_14': {'rawing': 'vitamin b12', 'level': 1}}
```

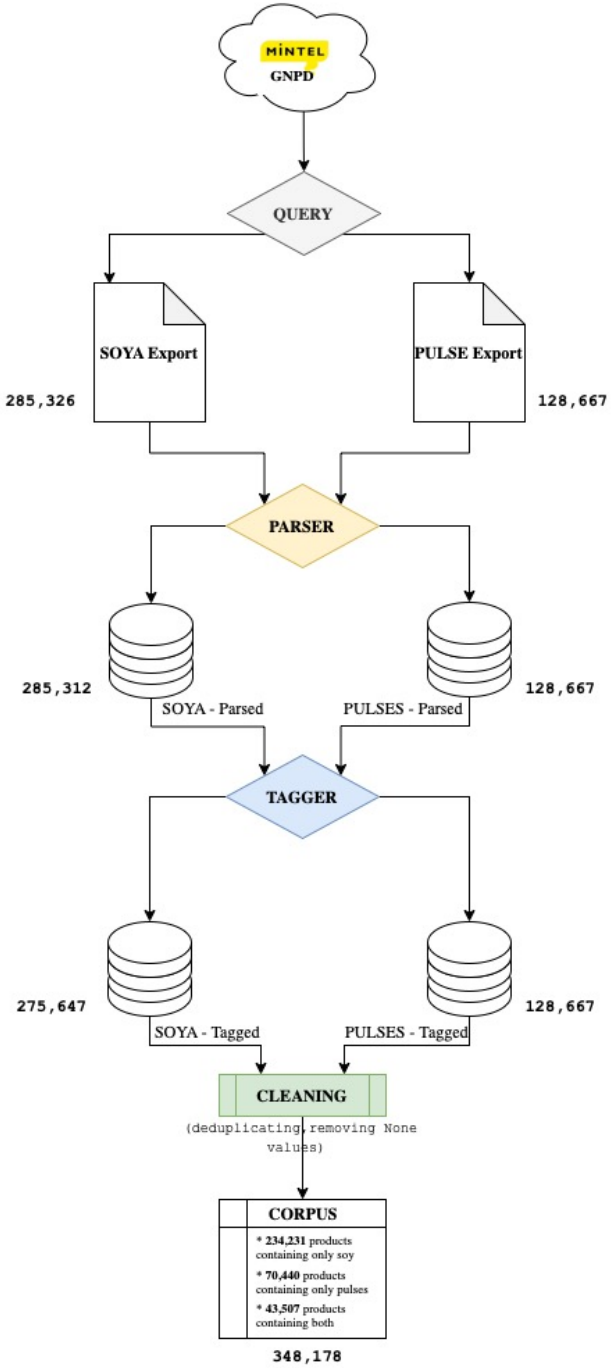
Excerpt 6. Dictionary of an ingredient list of multiple level.

In Salord et al., 2022

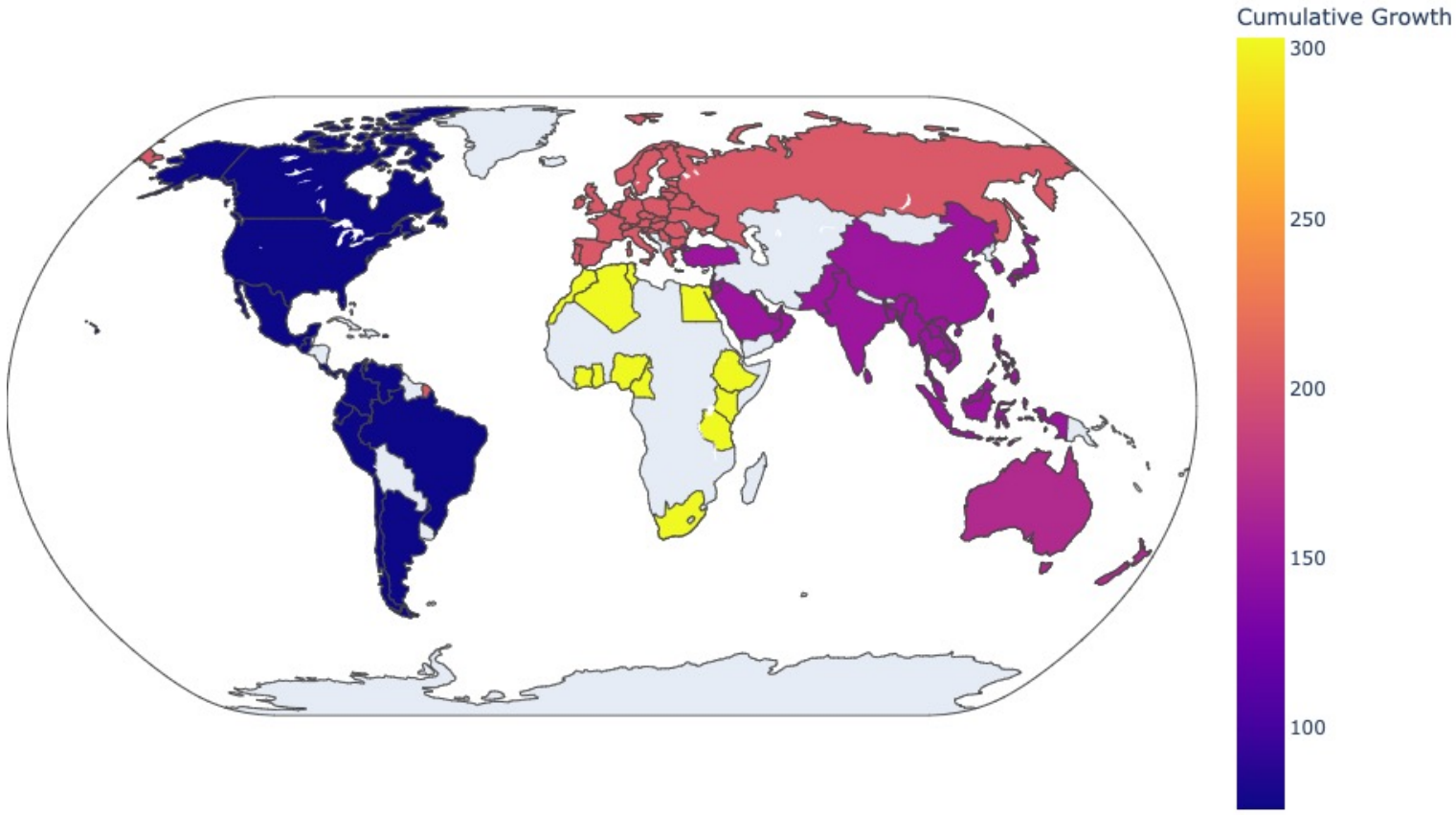
Données MINTEL (2010-2020) parsées

- Base de données GNPD-MINTEL collecte les « lancements produits »
  - réseau de collecte des innovations produits à l'échelle mondiale (n>86)
- Extraction de toutes les innovations produits contenant des espèces de légumineuses sur la dernière décennie, à l'échelle monde.
  - **T.A.L des liste d'ingrédients** : nettoyage, **parsing**, repérage des espèces en fonction de leur rang d'apparition, évaluation de leur impact sur le produit,...
  - **Base finale avec près de 300 000 produits** dont plus de 40 000 contenant soya et pulses



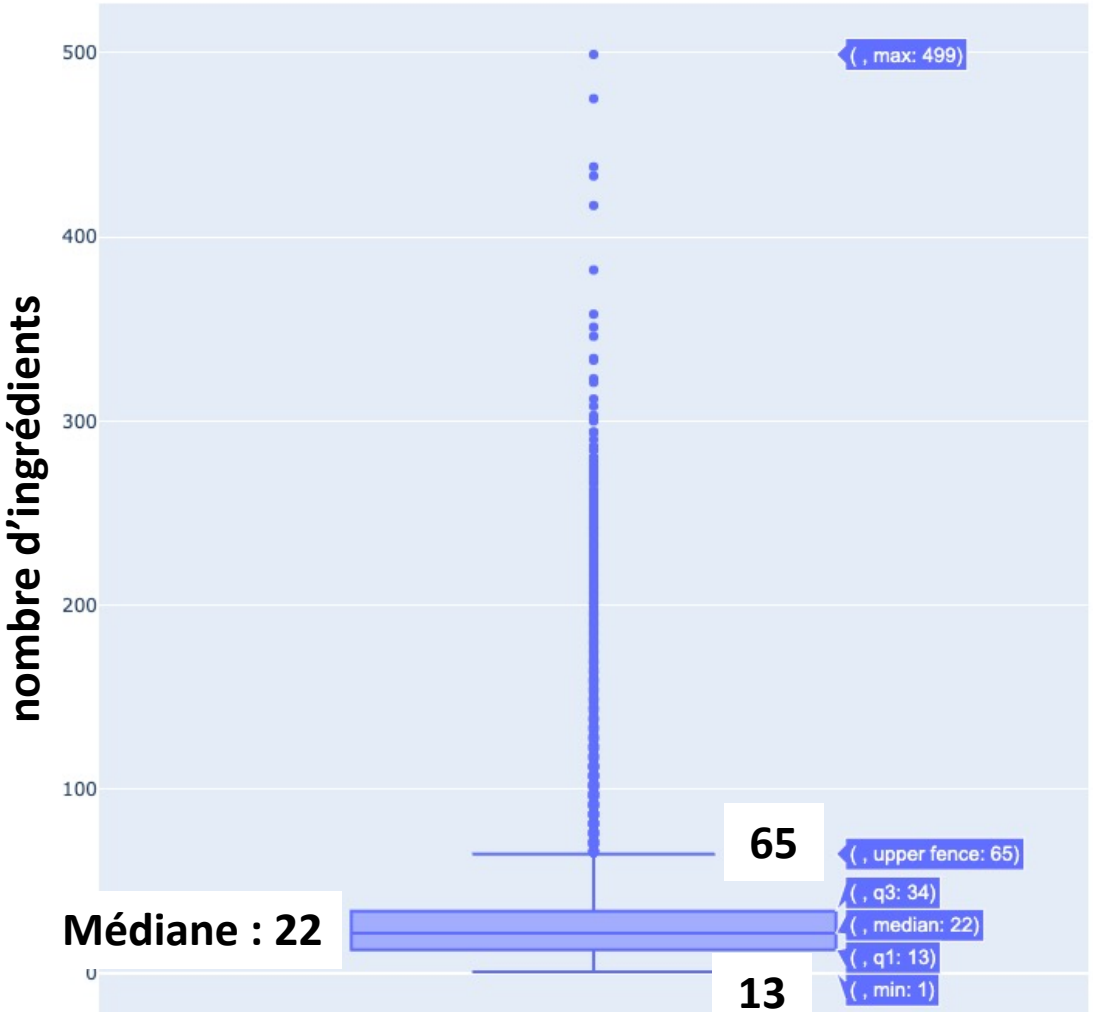


Carte établie à partir du taux de croissance moyen cumulatif du nb de lancements produits



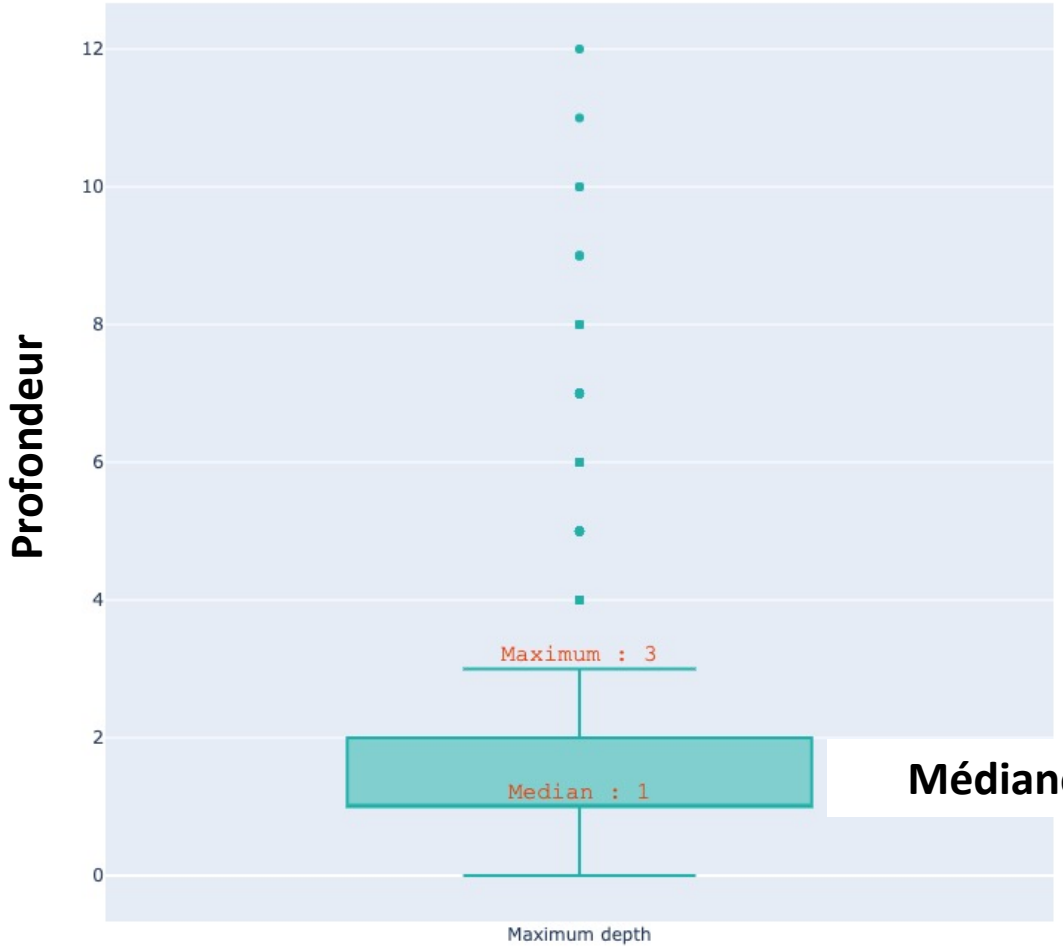
Nombre d’ingrédients et d’emboitement de listes d’ingrédients : ultra-transformation ou complexité ?

Distribution selon le nombre d’ingrédients



Médiane : 22

Distribution selon le nombre de listes « emboîtées»



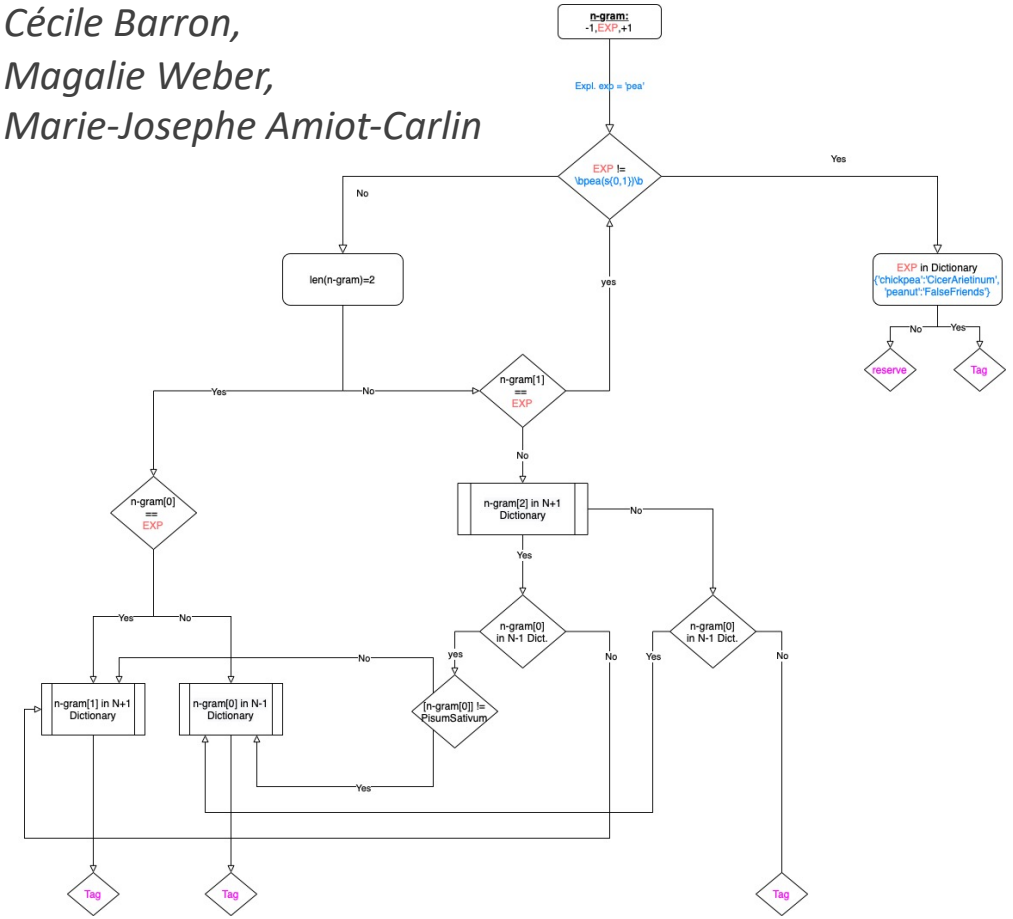
Médiane : 1

Identification des espèces de légumineuses à partir des listes d'ingrédients décomposées

→ **Une méthode d'analyse mixte mêlant automatisation et expertise professionnelle**

Comité d'experts FS&T (départements TRANSFORM et ALIMH):

- *Valérie Lullien-Pellerin,*
- *Cécile Barron,*
- *Magalie Weber,*
- *Marie-Josephe Amiot-Carlin*



Tx de croissance annuel cumulatif

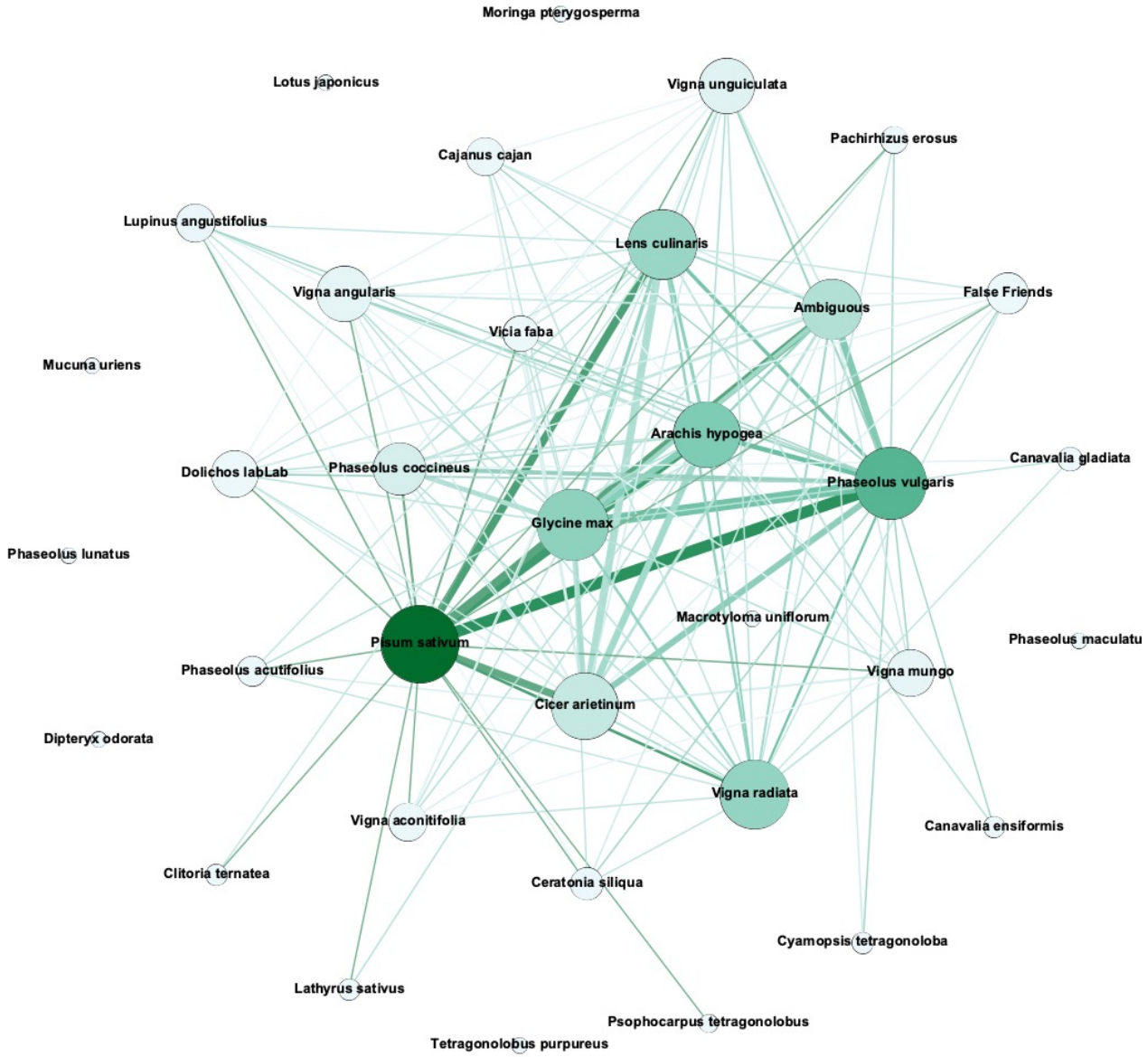
Nb de produits contenant l'espèce

Nb de produits contenant l'espèce parmi les 5 premiers ingrédients

Species	Raw Frequencies	Cumulative Growth (raw frequ.)	Presence in top 5 ing. Frequencies	Cumulative Growth (top 5 ing.)
Glycine max	277656	113.25	93359	116.88
Pisum sativum	36144	297.75	16291	302.00
Phaseolus vulgaris	25358	156.88	14668	141.25
Arachis hypogea	19324	144.00	12021	156.62
Cicer arietinum	16160	744.00	11560	823.50
Ambiguous	10862	132.50	5540	178.62
Lens culinaris	9706	523.00	7083	526.50
Ceratonia siliqua	7212	147.62	89	-0.00
Phaseolus coccineus	4601	211.75	2398	172.75
Vigna radiata	3899	413.50	2005	446.25
Lupinus angustifolius	1715	67.31	669	326.25
Vigna angularis	1284	466.75	698	275.00
Dolichos labLab	1162	62.91	447	7.41
Vicia faba	766	454.50	471	393.25
Vigna unguiculata	672	772.50	345	850.00

Le jeu des associations de légumineuses

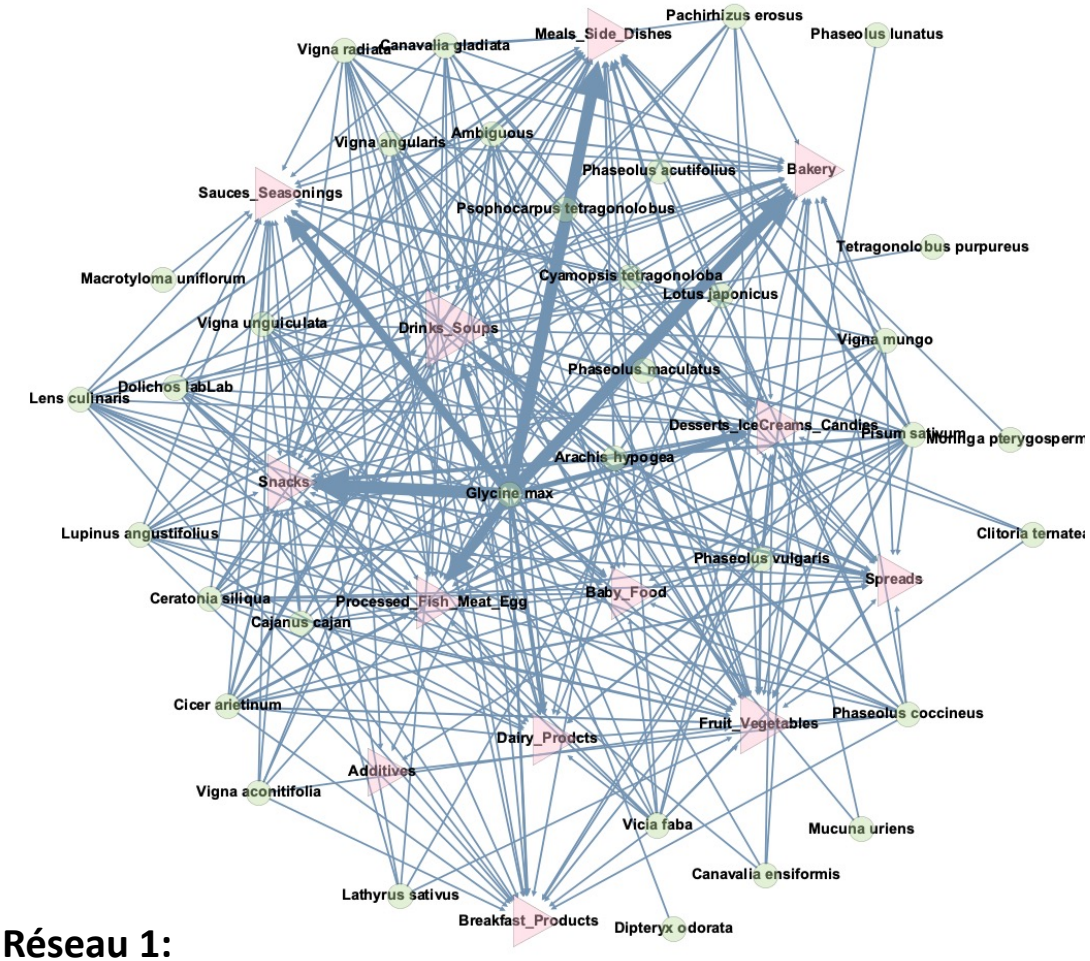
Exemple : le réseau de co-occurrence des espèces de légumineuses parmi les 5 premiers ingrédients



- **Ex:** Pois et haricot sont très souvent co-présents dans les innovations produits contenant au moins deux légumineuses parmi leurs premiers ingrédients.
- De façon générale lorsque plusieurs légumineuses sont associées il y a une forte probabilité qu'une de ces légumineuses soit du pois.

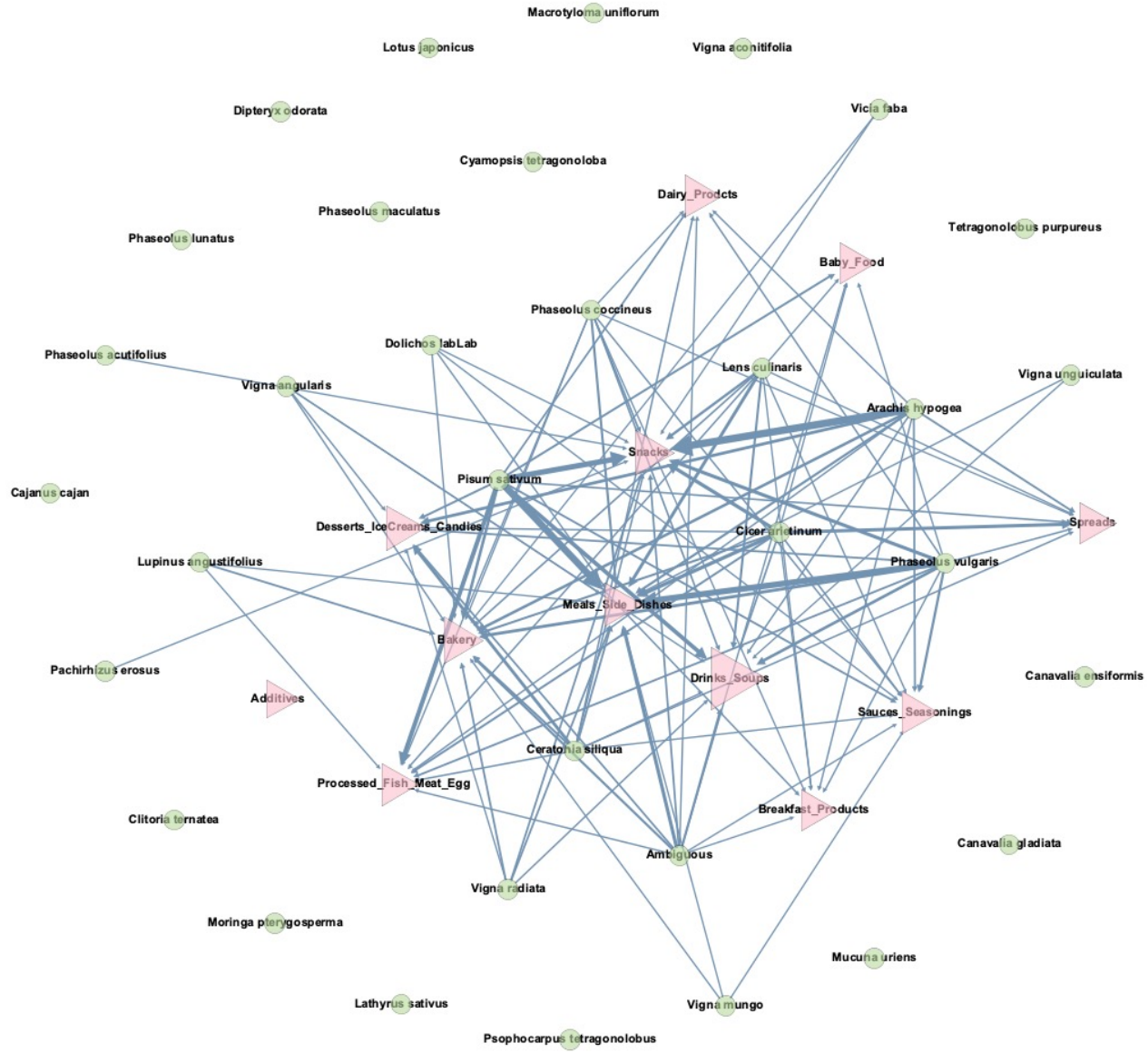
Spatialisation = Fruchterman Reingold,
Couleur = betweness centrality (https://fr.wikipedia.org/wiki/Centralité_intermédiaire).
La taille des nœuds est proportionnelle au degré sortant (les légumineuses citant le plus d'autres légumineuses).
L'épaisseur des liens représentent le nombre de fois qu'un couple de légumineuse est cité ensemble.

Association des ingrédients et segments de produits



Réseau 1:
Réseau comprenant le soja.

Réseau 2:
Réseau en enlevant le soja et filtrant les liens par leur poids (occurrence supérieure ou égale à 100)



Spatialisation = Fruchterman Reingold,
Couleur = type d'entité . N La taille des nœuds est proportionnelle au degré entrant (les entités étant le plus « cités » par d'autres). L'épaisseur des liens représentent le nombre de fois qu'un couple de légumineuse/type de produit est identifié.

Prochaine étape : catégorisation des ingrédients légumineuses au regard de la transformation subie & contribution aux ontologies

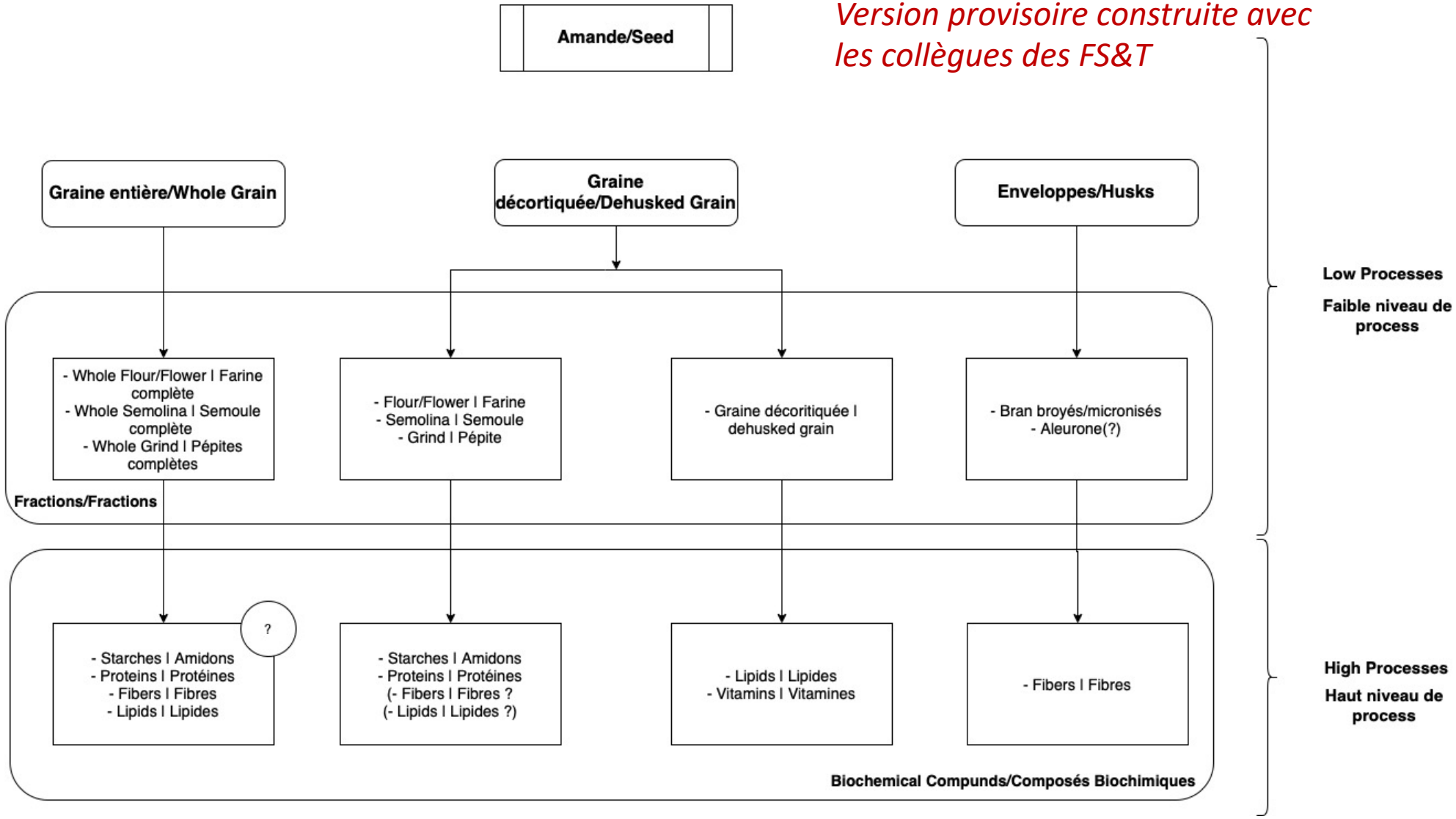
Catégoriser les usages des espèces au sein des listes d'ingrédients,

Vers un **appariement avec les principales ontologies** agro-alimentaires,

Analyses produits profils ingrédients/labels, profils ingrédients/allégation marketing, … par **ACM**

→ Comment prendre en compte les types d'ingrédients dans la définition d'un produit ultra-transformé ?

Version provisoire construite avec les collègues des FS&T



Conclusion

- **L'innovation est une clé de sortie du déverrouillage sur les légumineuses, mais quelles sont trajectoires d'innovation qui asseoient leur contribution à des systèmes alimentaires sains et durables ?**
  - Maintien suprématie du soja ou vers une plus grande diversité de légumineuses ?
  - Développement d'une offre ultra-processée pour quelles fonctionnalités, usages ?
  - ….

  L'analyse des ingrédients, des formulations et autres mentions reste essentielle pour objectiver les *narratives* associés aux voies de développement, aux modèles de l'agroalimentaire et leurs transformations

- **Couplage aux données de consommation : quelle offre est privilégiée par le consommateur ?**
  - Quelles difficultés de positionnement (image, prix…) d'une offre « moins processée » par rapport à une offre « plus processée » ?