

INRAE

Département
MathNum

MaiAGE

➤ Des textes aux connaissances avec les LLM : extraction, structuration et fiabilité

Claire Nédellec et l'équipe Bibliome

Modèles de fondation pour l'extraction d'information structurée à partir de texte dans un contexte FAIR

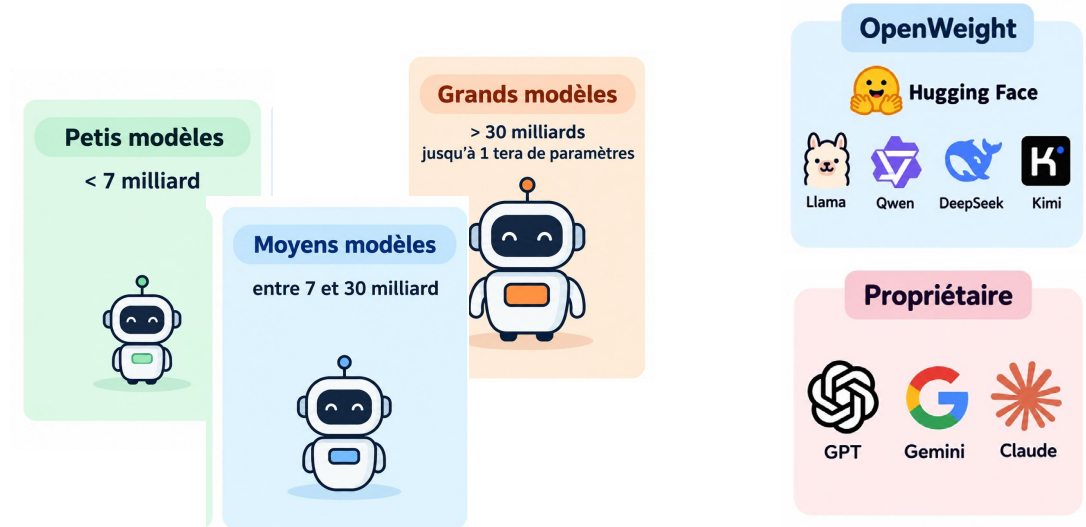
Les grands modèles de langage (LLM)

Modèles de fondation spécialisés dans la compréhension et la génération en langage naturel multilingue (*Natural Language Processing*) et en langage informatique (code).

Aujourd'hui s'étendent au multimodal, OCR, raisonnement

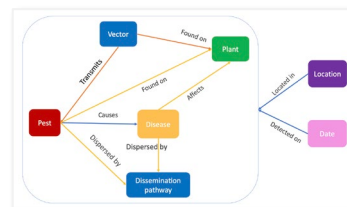
Extraire, structurer et normaliser

automatiquement les informations à partir de texte en fonction d'un schéma prédéfini pour les agréger entre elles et les comparer avec les informations connues



Exemple sur les relations écologiques

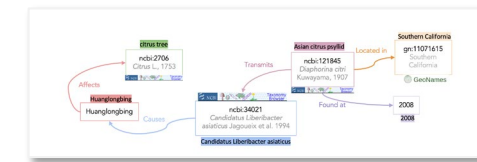
Interactions biologiques
Ancrage spatiotemporel



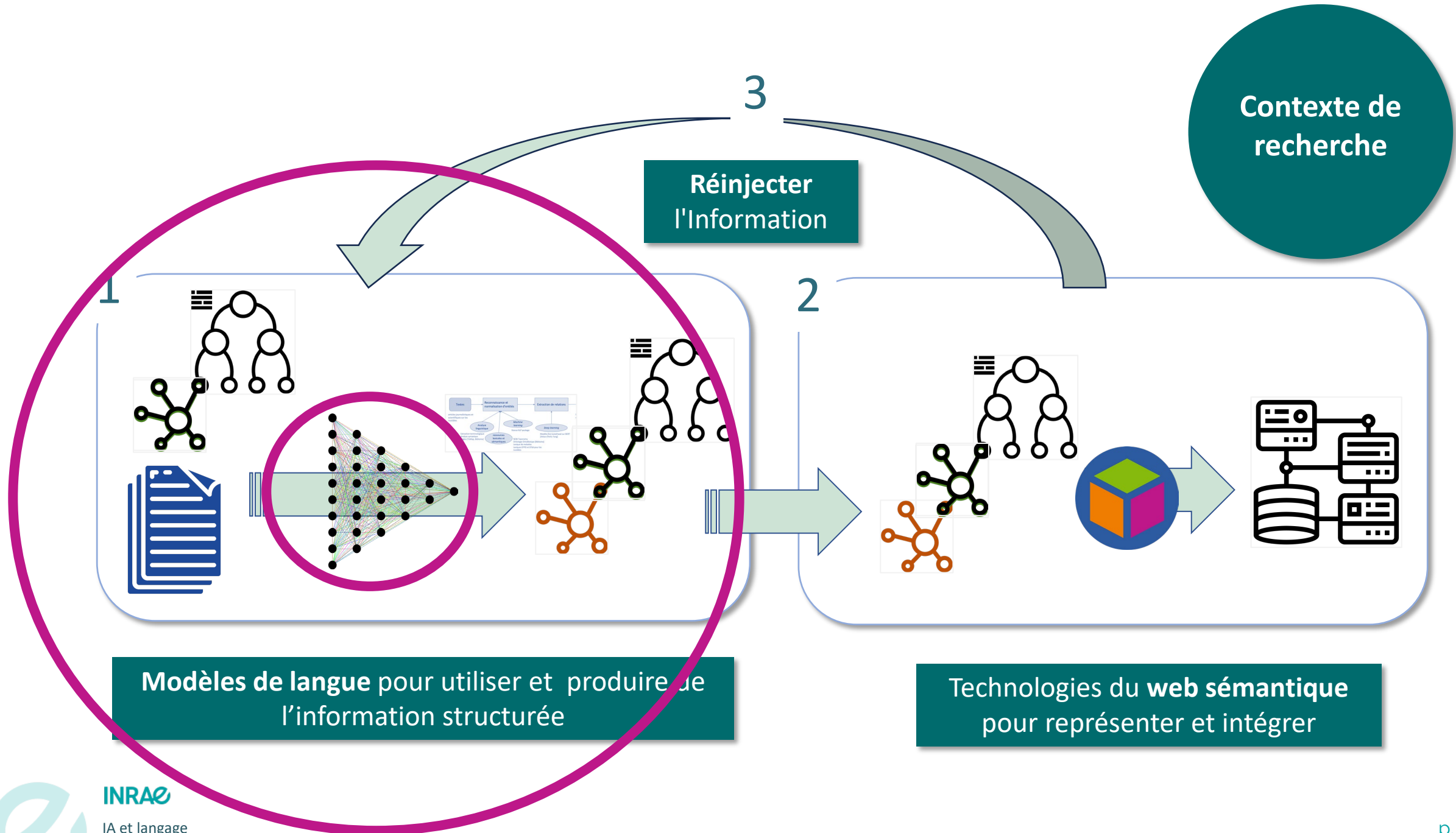
Ontologie et références

Beyond feeding damage, ACP vectors the lethal bacterium (*Candidatus Liberibacter asiaticus*) that causes Huanglongbing (HLB or citrus greening) disease. Huanglongbing causes leaves to yellow asymmetrically (blotchy mottle), and results in unmarketable fruit with asymmetrical shape, aborted seeds, and bitter flavor. The disease can kill a citrus tree within 5 to 8 years, and there is no known cure for the disease. Asian citrus psyllid arrived in Southern California from Mexico in 2008 and is now well established in Southern California and spreading northward. Huanglongbing was first found in residential trees in Los Angeles County in 2012, likely due to illegal importation of infected plant material; it has spread from that area and also spread northward from Mexico.

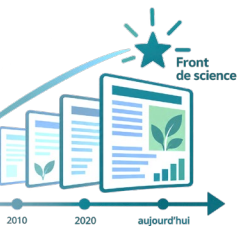
Informations extraites du texte



Graph de connaissance normalisé



L'équipe Bibliome de MaIAGE



Equipe de recherche en Traitement automatique de la langue depuis 20 ans

Forte évolution en 10 ans : *deep learning*, LLM

Robert
Bossy
IR



Arnaud
Ferré
CR



Claire DR
Nédellec



thèse
Xingyu
Zhu



thèse
Anne-
Sophie
Foussat



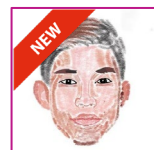
CR
Louise
Deléger



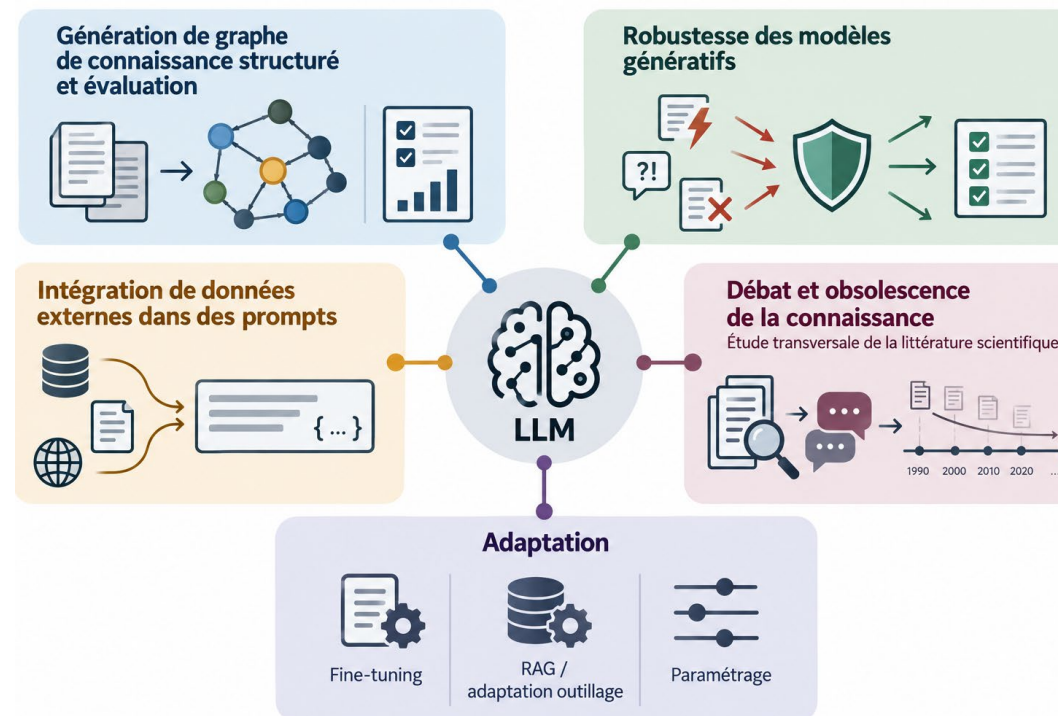
Marine
Courtin
Post-doc



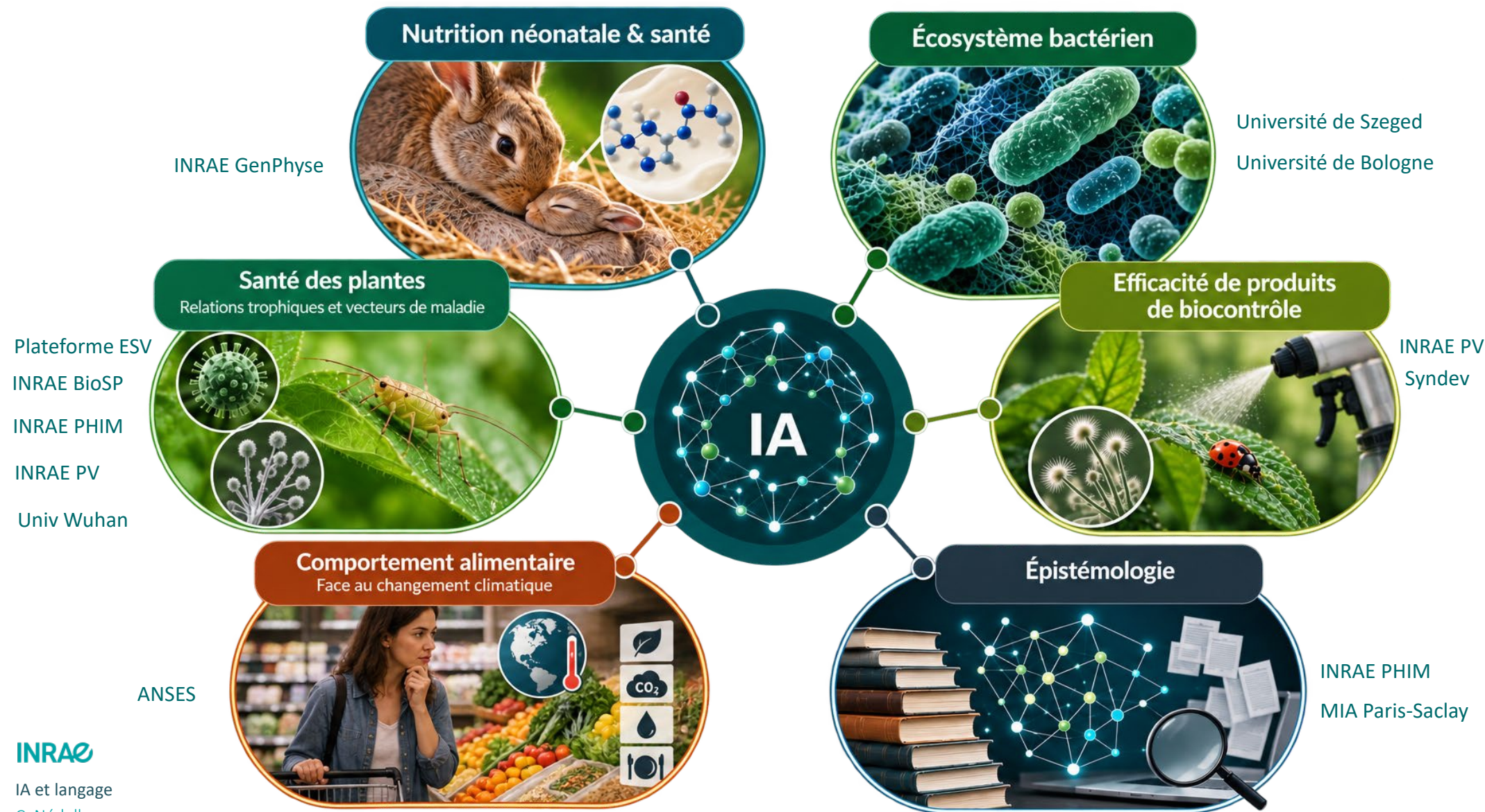
IR
Mathieu
Lai King



Sélection de travaux Focus LLM



Une recherche interdisciplinaire



Génération de graphe de connaissance structuré à partir de textes *pour la surveillance de la santé des plantes*

Interactions
biologiques
Temps
Espace

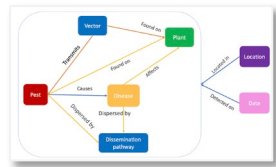
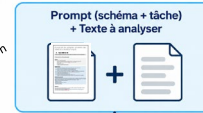
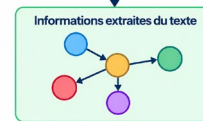


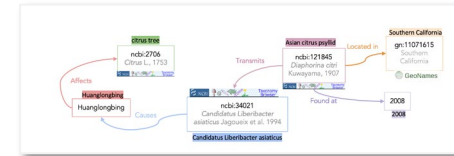
Schéma de données

Traduire le schéma en instruction



Traduire le texte généré en graphe

Beyond feeding damage, **ACP** vectors the lethal bacterium (*Candidatus Liberibacter asiaticus*) that causes **Huanglongbing** (HLB or citrus greening) disease. **Huanglongbing** causes leaves to yellow asymmetrically (blotchy mottle), and results in unmarketable fruit with asymmetrical shape, aborted seeds, and bitter flavor. The disease can kill a **citrus tree** within 5 to 8 years, and there is no known cure for the disease. **Asian citrus psyllid** arrived in Southern California from Mexico in 2008 and is now well established in Southern California and spreading northward. **Huanglongbing** was first found in **residential trees** in Los Angeles County in 2012, likely due to illegal importation of infected plant material; it has spread from that area and also spread northward from Mexico.



Graphe de connaissance normalisé

Défis de recherche



Evaluation des méthodes de génération de textes

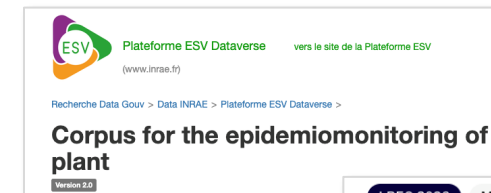
Conception du corpus EPOP (*Epidemiomonitoring of Plants*), corpus de référence de qualité

Engagement de communautés interdisciplinaires

- Animation d'un groupe d'experts sur les bioagresseurs (32 personnes)
- Organisation du challenge international **PestCLEF** dans le LifeClef Lab de *CLEF Conference and Labs of the Evaluation Forum* (12 équipes)

Méthodes basées sur les LLM : ensembles, prompting. Performances de **DeepSeekV3** : 72% F-mesure

Défis : analyse du discours et inférence



INRAE

IA et langage
C. Nédellec

Projets **anr**® BEYOND (PPR CPA) et EcoControl (PEPR AgriEcoNum)
Tiers-ESV (IB), KPREL (DataIA, CSC)

Robustesse des LLM pour l'extraction d'information structurée

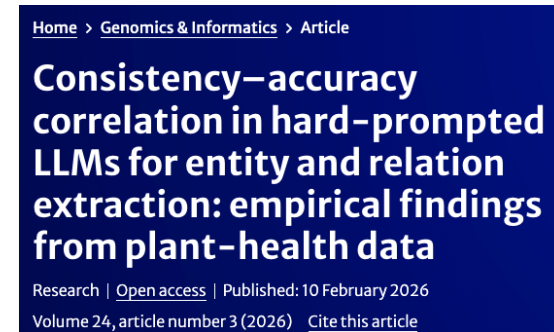
La cohérence et l'exactitude des prédictions en réponse à une même question sont-elles corrélées ? Non, faiblement

- La stabilité des réponses dépend du **modèle** et de la **complexité de la tâche**. (modèles GPT, DeepSeek, Qwen et Kimi sur le corpus EPOP)
- Concl: Exactitude et cohérence des modèles ne doivent pas être évaluée au niveau textuel, mais sémantique en raison de la variabilité des formes

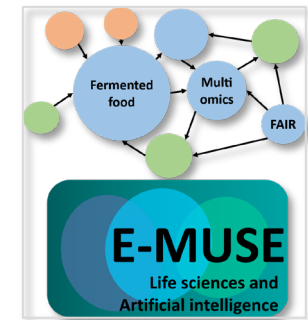
Redéfinir les protocoles d'évaluation

La méthode SCoNE : apprentissage bruité et auto-correction (modèle T5)

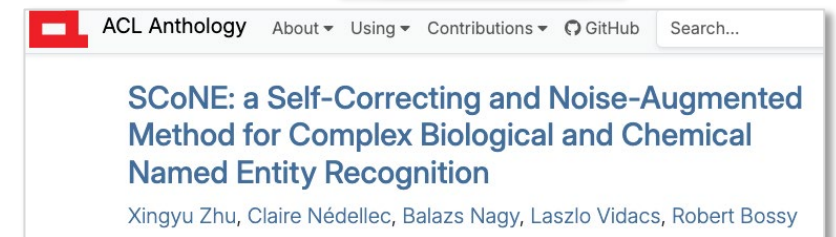
- **Entraînement** : bruitage par masquage progressif des entités de référence → reconstruction → apprentissage robuste des structures complexes
- **Inférence** : quand la confiance du modèle est faible, au niveau token, alors régénération ciblée
- **Résultat pour la reconnaissance d'entités complexes** : +1,8 à +2,7 points de F1 sur les benchmarks CHEMDNER et Florilège (fermentation microbienne).



Collaboration avec HZAU, univ d'Agriculture de Wuhan



Collaboration avec univ Szeged
Projet DTN FairOmics

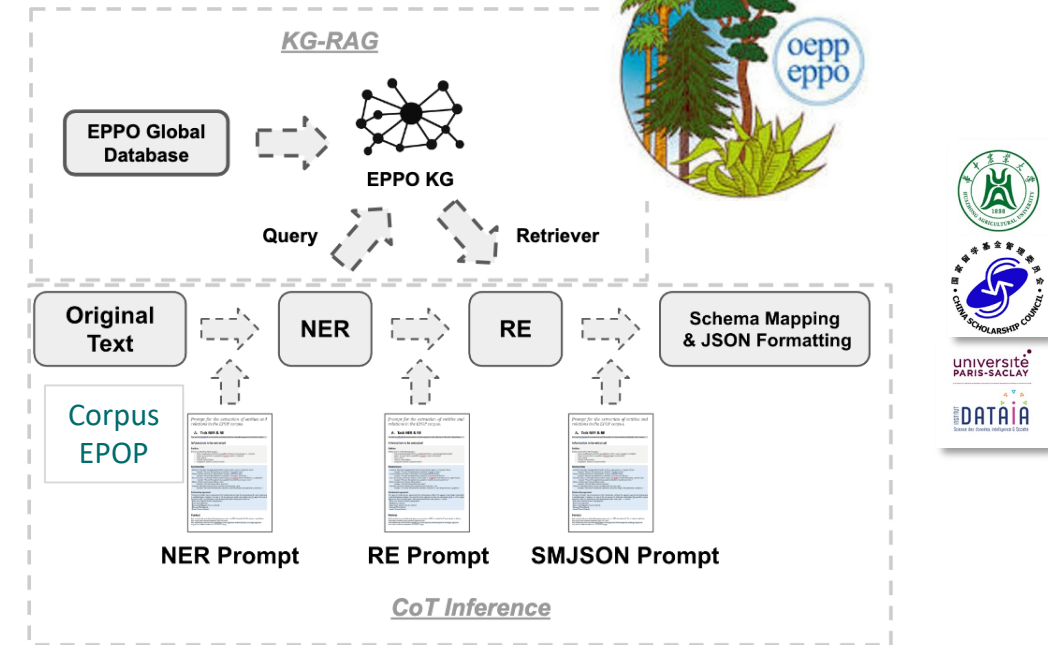


Formuler des instructions efficaces

Knowledge Base-Enhanced Prompting for Relationship Extraction using Large Language Models (KPREL)

- Choix de l'information externe pour guider le LLM dans l'extraction de relation : ajout des chemins du graphe de connaissance de longueur $< k$ dont les entités sont similaires
- Générer la réponse structurée en langage naturel
- Conformité au schéma d'extraction : *Schema-Matching JSON*

Model	Precision	Recall	F1-score
Qwen3-32B (thinking)	0.4	0.4	0.38
Qwen3-32B + COT + RAG + SMJSON	0.857	0.462	0.571



RAG Switch — Quand faut-il chercher de l'information externe ?

Ne déclencher le RAG que lorsqu'il est utile.

- **Décision adaptative** : estimation de la difficulté de prédiction pour chaque cas par EL2N et des signaux d'incertitude.
- **Apprentissage** : le besoin de RAG est déterminé à partir du **succès/échec de l'extraction sans RAG**.
- **Résultat** : meilleur compromis performance / coût pour l'extraction d'information



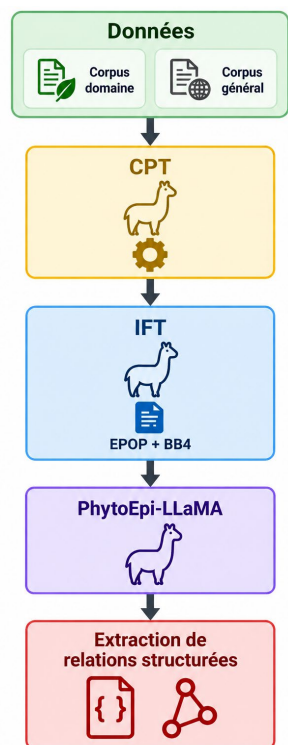
Task-Grounded and EL2N-Informed RAG Switch for Domain-Specific Information Extraction

Xingyu Zhu^{1,2}, Robert Bossy¹, Balazs Nagy², Laszlo Vidacs², Claire Nédellec¹

¹INRAE, Université Paris-Saclay, Paris, France

²Department of Software Engineering, University of Szeged, Szeged, Hungary

Garantir des sorties structurées et conformes avec des modèles LLM plus sobres, plus petits et adaptés au domaine



Adapter un « petit » LLM au domaine : *PhytoEpi-LLaMA*

pré-entraînement continu : CPT (*Continual Pre-Training*) améliore la maîtrise du langage du domaine par le modèle, mais peut dégrader la qualité de la structure de la réponse.

Ajustement sur instructions : IFT (*Instruction Fine-Tuning*) apprend la tâche particulière et restaure la conformité des sorties.

Evaluation sur le corpus EPOP : Llama-3.1-8B adapté (+CPT+IFT avec les tâches EPOP +BB4) : proche des performances de modèles 32B en *prompting* pour un coût réduit.

Conclusion. Adaptation au domaine = langage du domaine + apprentissage de la tâche et respect du schéma

PhytoEpi-LLaMA: Lightweight Domain Adaptation for Reliable Information Extraction in Plant Health Literature

Xinzhi Yao¹, Claire Nédellec¹, and Robert Bossy¹

Université Paris-Saclay, INRAE, MaIAGE, 78352 Jouy-en-Josas, France



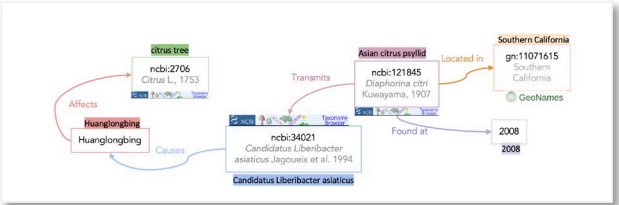
Association for
Computational Linguistics



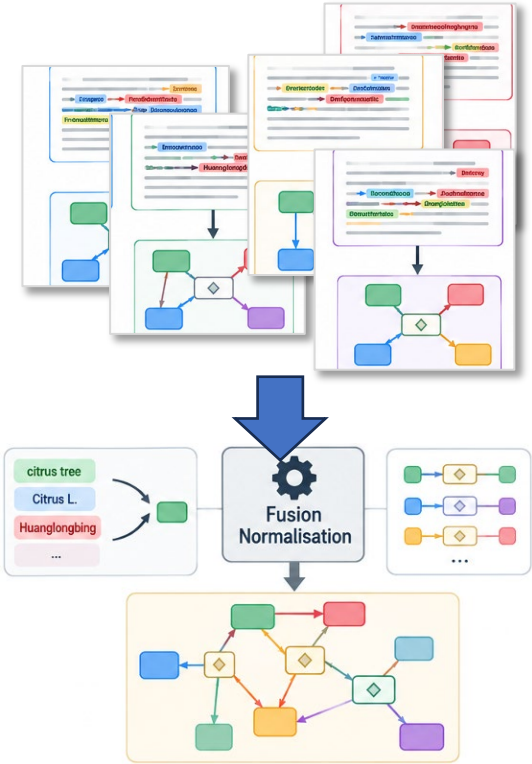
Soumis

Du texte au graphe de connaissances, consolidation et prise en compte du débat

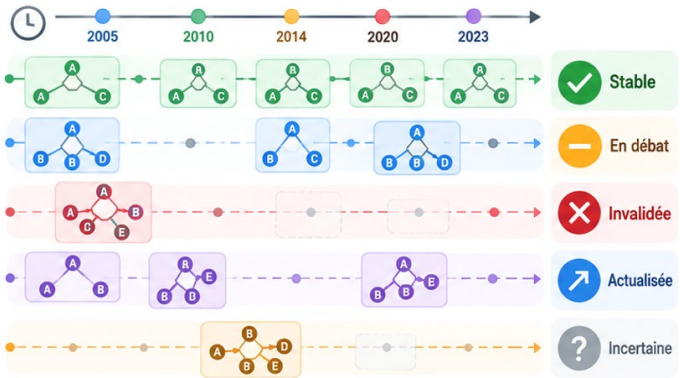
Beyond feeding damage, **ACP** vectors the lethal bacterium (**Candidatus Liberibacter asiaticus**) that causes **Huanglongbing** (HLB or **citrus greening**) disease. **Huanglongbing** causes leaves to yellow asymmetrically (blotchy mottle), and results in unmarketable fruit with asymmetrical shape, aborted seeds, and bitter flavor. The disease can kill a **citrus tree** within 5 to 8 years, and there is no known cure for the disease. **Asian citrus psyllid** arrived in Southern California from Mexico in 2008 and is now well established in Southern California and spreading northward. **Huanglongbing** was first found in residential trees in Los Angeles County in 2012, likely due to illegal importation of infected plant material; it has spread from that area and also spread northward from Mexico.



Extraire de façon structurée et normalisée

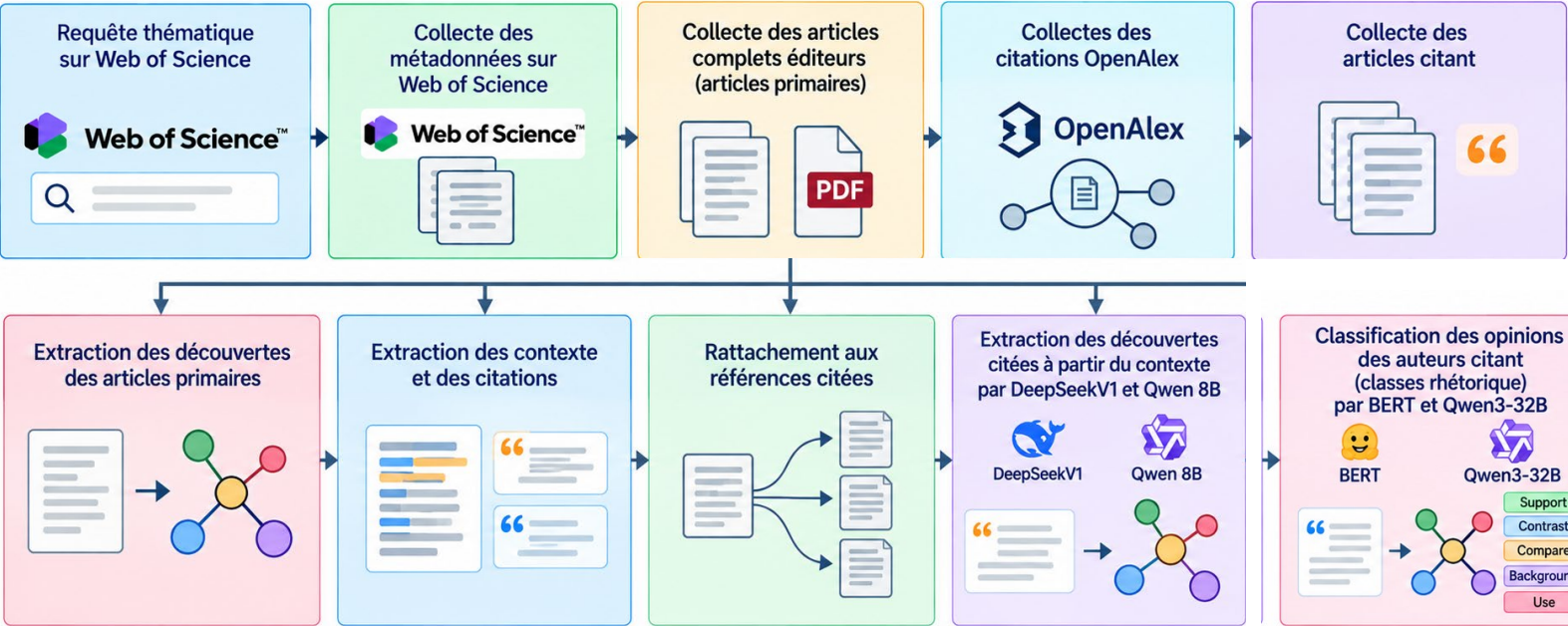


Comparer et agréger pour construire/enrichir les graphes de connaissance :



Prendre l'échelle temporelle en compte pour estimer la fiabilité

Estimer la fiabilité par l'analyse des opinions des pairs dans le contexte des citations - application à la transmission vectorielle des maladies

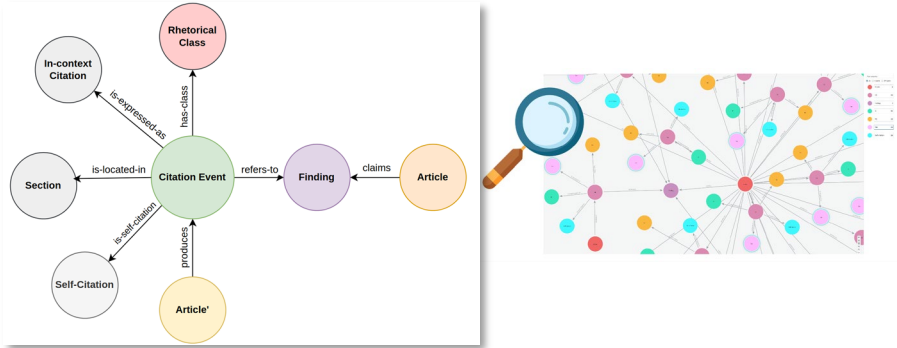
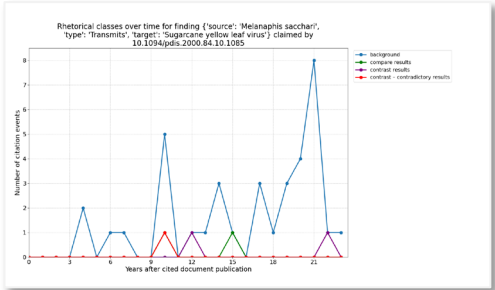


Collection	Count
WoS primary references	8,181
Query Article Set (with DOI)	7,318
Citing references	94,042
Union	97,800
Full texts retrieved	62,878

Corpus de référence
PD100Cit : classes rhétoriques annotées
Informations EPPO disqualifiées: article associé à une découverte niée ou douteuse.

Class	Precision	Recall	F1-score	Support
Compare	0.958	0.862	0.907	289
Contrast	0.664	0.798	0.725	89
Macro average	0.811	0.830	0.816	378
Weighted average	0.888	0.847	0.864	378

Graph element	Instances
Articles (primary and citing)	87,974
Original findings	6,969
Citation events (CEs)	262,807
Characterized CEs	169,925



Extraire les informations citées, les comparer, dégager un avis

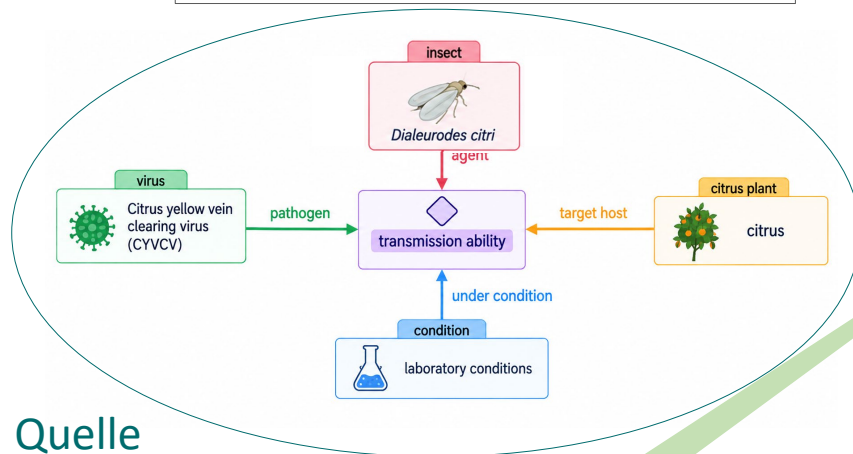
Article de (Zhang et al., 2019)

This is the first report of the ability of *D. citri* to transmit **CYVCV** from infected to **healthy citrus** under **laboratory conditions**

Article de (Bin et al., 2026)

The rapid spread of CYVCV in **citrus orchards** [...]. Additionally, **laboratory studies** have shown that **CYVCV** can also be transmitted by insect vectors, including *Aphis spiraecola*, *A. craccivora*, and *Dialeurodes citri* (Onelge et al. 2010; Zhang et al. 2019a).

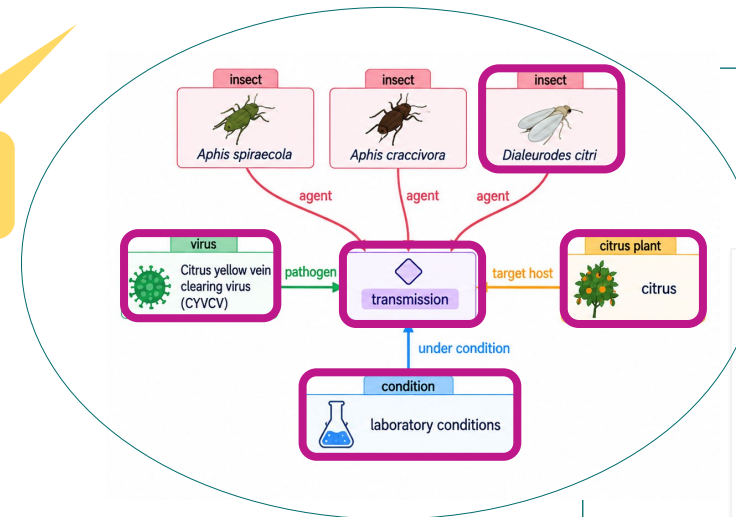
Classe
rhétorique
prédite
Background



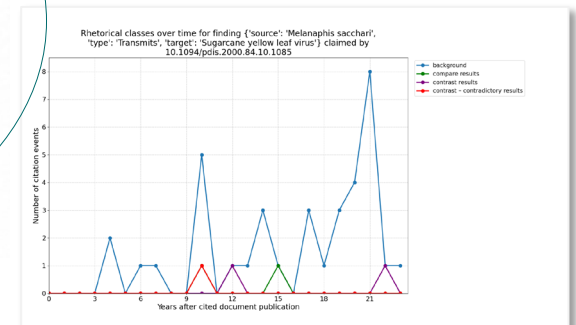
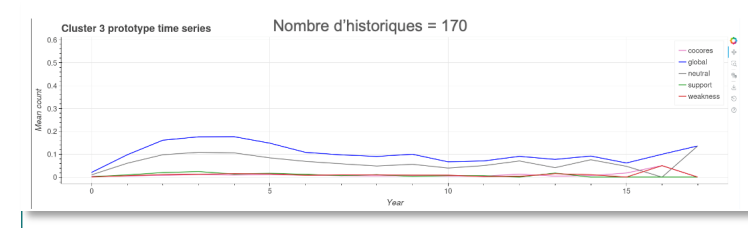
Quelle
valeur ?

Zhang et al.
2019a

Cité par Bin et
al. 2026



Cluster de séries temporelles
pour rapprocher l'instance
d'un profil caractéristique

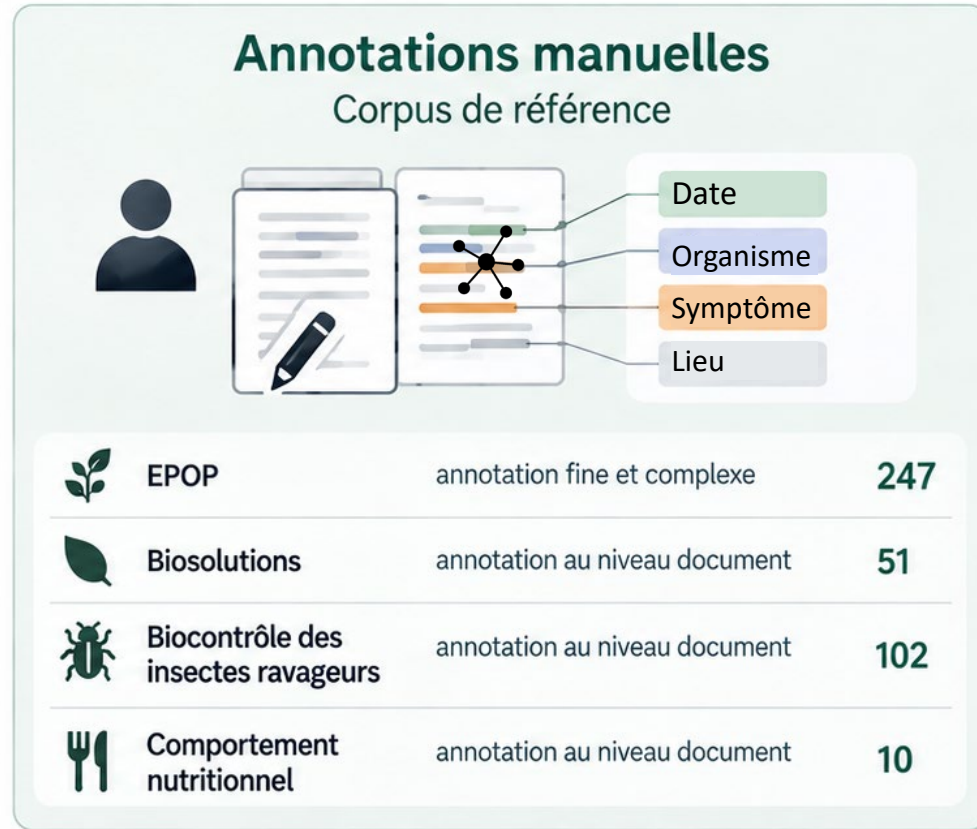


1. Extraire les résultats primaires
2. Relier citation et publication primaire
3. Extraire les résultats cités
4. Comparer le résultat primaire au résultat cité
5. Si égal, prédire la classe de l'opinion
6. Répéter et tracer le profil du résultat au cours du temps

INRAE

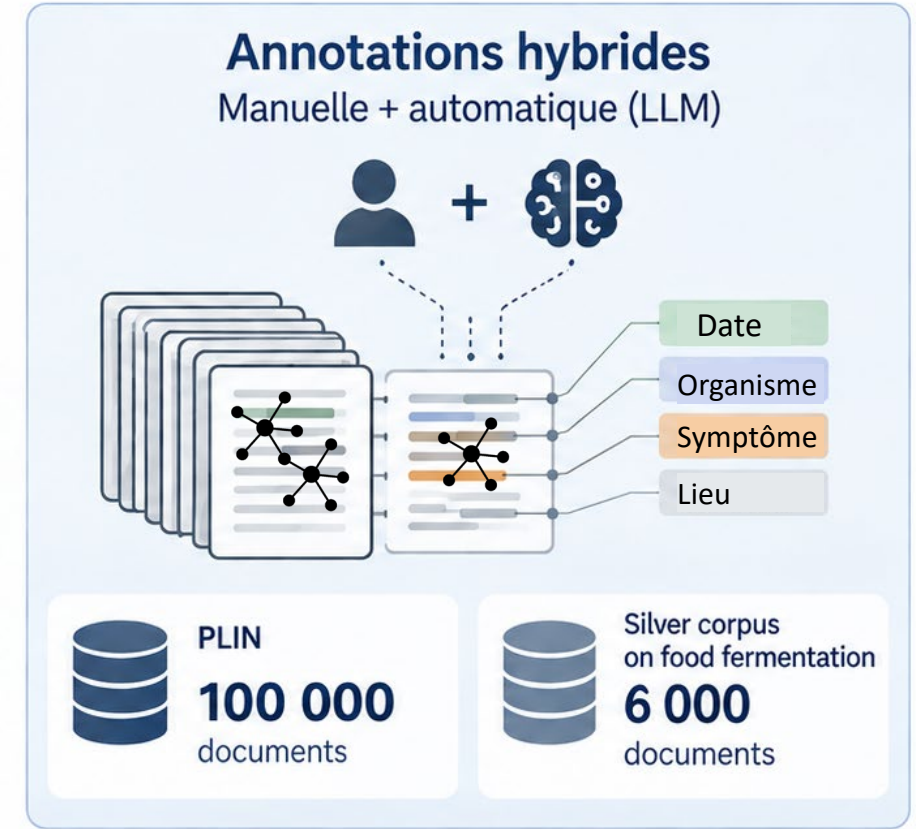
IA et langage
C. Nédellec

Entraîner des modèles et évaluer avec des annotations manuelles ou automatiques



Changement
d'échelle

De corpus experts
à des corpus massifs



Qualité de référence
(annotations manuelles, riches)

Passage à l'échelle

Supervision hybride
(manuelle + automatique)

Perspectives : utiliser les LLM pour construire, justifier et maintenir une connaissance scientifique fiable évolutive

1. Du graphe extrait au graphe consolidé

Fusionner les graphes issus de milliers de documents : normalisation, corréférence, redondance, contradictions, relations n-aires.

2. Adaptation

Sélection des modèles

Comment **adapter** ? CPT, IFT, prompting, RAG ou combinaison de ces approches

Avec quelles **données** ? Corpus experts, *silver data*, annotation hybride.

À quel **coût** ? Compromis performance × robustesse × coût de calcul.

3. De l'extraction adaptative à l'extraction ouverte

De schéma + prompt fixes à prompt / RAG dynamiques jusqu'à la découverte et l'évolution de schéma

4. De la performance à la confiance

Pas seulement « l'extraction est-elle correcte ? », mais « peut-on savoir quand elle est fiable ? » : incertitude, cohérence entre générations, calibration, auto-correction.

4. Un enjeu transversal : évaluer le système complet

Passer de métriques locales (F1) à une évaluation de bout en bout : texte → extraction → graphe → fusion → connaissance exploitable, avec traçabilité depuis les sources.

5. De l'extraction à la dynamique de la connaissance

Caractériser une connaissance comme confirmée, discutée, révisée ou obsolète par analyse longitudinale

INRAE

