# Méthode de construction de bases de connaissances à partir d'alignements entre ontologies de référence

Résumé : Le but général de cette thèse sera le développement de méthodes pour produire des bases de connaissances complexes qui intègrent des ontologies de référence. La difficulté consiste à réaliser des correspondances, non pas en aveugle, mais en prenant en compte les cas d'usage et les sources d'information associées. Pour ce faire nous définirons des patrons de conception ontologique dans le domaine agricole. La problématique principale de cette thèse est de produire des correspondances N-aires qui concernent à la fois des entités sémantiques et les entités des sources. Notre cadre applicatif sera l'évolution des risques d'attaque des cultures à partir des Bulletins de Santé du Végétal.

## Contexte et enjeux :

Les données sont publiées sur le Web à l'aide des technologies du Web Sémantique dans le but de simplifier la découverte de données et de répondre à la problématique d'intégration de données hétérogènes.Pour pouvoir intégrer plusieurs jeux de données agricoles, les mettre dans un format informatique interopérable ne suffit pas. Dans un souci d'intégration efficace, des jeux de données distincts doivent être liés entre eux par le biais de données pivot.Des ontologies de référence définissant des données pivot commencent à être publiées sur le Linked Open Data (LOD) ou en français Web de données liées [Bizer et al, 2009] [Jonquet et al, 2015]. Elles sont néanmoins loin d'être couvrantes et nous sommes toujours amené à construire de nouvelles bases de connaissances qui permettent de couvrir nos cas d'usage spécifiques. Le but général de cette thèse sera le développement de méthodes pour produire des bases de connaissances complexes qui intègrent des ontologies de référence en agriculture. La difficulté consiste à réaliser des correspondances, non pas en aveugle, mais en prenant en compte les cas d'usage et les sources d'information associées. Nous nous retrouvons donc avec une problématique de mise en correspondance N-aire qui concerne à la fois des entités sémantiques et les entités des sources.

## Cas d'usage agricole

L'équipe COPAIN a travaillée sur la collecte d'un corpus de bulletins d'information agricole intitulé Bulletin de Santé du Végétal (BSV). Ces documents ont été annotés pour faciliter la recherche en fonction de certains critères (culture, lieu, période temporelle). Le corpus et ses annotations spatio-temporelles ont été publiés sur le LOD [Roussey et al, 2016]. Les annotations spatio-temporelles produites ne sont pas suffisantes pour répondre à tous les besoins d'analyse. D'autres ontologies de références sont nécessaires. Nous avons encore besoin de construire des bases de connaissances pour décrire :

- Les principaux agresseurs des cultures cultivées en France.

- Les observations des stades de développements des cultures
- Les niveaux de risque des attaques d'une culture.

Par exemple, la base de connaissance sur les agresseurs des cultures devrait être liée au référentiel des cultures. Il s'agit d'établir des relations binaires entre un agresseur et une culture. La FAO a défini dans son ontologie la propriété "pestOf" pour représenter cette relation binaire [AOS].

L'observation d'une attaque d'une culture par un agresseur est une relation N-aire. Cette relation lit une culture ayant atteint un stade de développement à un agresseur donné. Une date d'observation ainsi que des informations spatiales sur la parcelle cultivée sont aussi indiquées.

La base de connaissances des agresseurs des cultures devrait indiquer le risque à partir du stade de développement des cultures et des observations d'attaque. L'évaluation du risque n'est pas aisée; en effet, les risques identifiés dans les BSV ne sont valables que si la culture a atteint un certain stade de développement. Il s'agit de reconnaître des relations N-aires entre une culture, un stade de développement et un agresseur. Chacun des noeuds doit venir d'une ontologie de référence. Par exemple le code G2 est défini dans l'ontologie associée comme le stade de développement correspondant à la formation des premières siliques. Il faudra aussi mettre en place un processus de raisonnement pour la validation et la complétion de ces relations.

# Objectifs de la thèse

L'objectif de cette thèse est de proposer une nouvelle méthode d'acquisition de relations N-aires entre éléments de plusieurs ontologies, qui exploite des corpus de texte. Ces travaux sont une continuité de la thèse de Fabien Amarger [Amarger 2015].Notre but sera de bénéficier des outils et méthodes proposés par un laboratoire en informatique pour les adapter aux problématiques de la construction d'ontologies agricoles.Le LIPN propose une méthode pour déterminer des relations binaires à partir de documents [Ben Abbes, 2013]. Cette thèse proposera une nouvelle méthode étendant les approches existantes pour produire des relations N-aires dirigées par des patrons de conception du domaine agricole. L'acquisition à partir de textes, qui sera fournit par le LIPN, permettra d'alléger l'intervention des experts spécialistes et les aide dans le processus de conception et de formalisation des connaissances (les relations entre les entités sémantiques des ontologies) [Szulman *et al*.,2009].

L'ancrage à des ontologies existantes permet à la fois de réutiliser des référentiels établis et de créer des liens qui viennent enrichir l'existant. Ce travail participera à l'effort de mise en commun et de partage de connaissances porté par la communauté du Web sémantique.

Cette thèse s'intéressera à développer des patrons de conception ontologiques dans le domaine de l'agriculture ainsi que des anti-patrons. En effet, la production automatique à partir de documents (au contraire de la transformation de sources structurées) ne permet pas d'établir de manière certaine l'existence d'une relation entre N entités issues d'ontologies différentes. Les patrons de conceptions permettent de diriger la méthode d'acquisition des relations N-aires. Les anti-patrons permettent de nettoyer les candidats de relations N-aires quand l'existence de ces relations mettent en évidence des incohérences au sein de la base de connaissances.

Les défis sont multiples. Nous travaillons sur des patrons du domaine agricole [Roussey et al, 2013]. La FAO par le biais de son ontologie propose aussi des patrons dans le domaine de l'agriculture [Fiorelli et al, 2014]. Les patrons de la FAO sont des relations binaires, nous devrons les enrichir pour produire des patrons pour des relations N-aires.

- Objectif 1: Proposer ou étendre des patrons de conception ontologiques dans le domaine de l'observation des cultures et de leur protection.
- Objectif 2: Déterminer une méthode d'identification de relations N-aires dirigée par des patrons du domaine et tenant compte de l'incomplétude d'une relation
- Objectif 3: déterminer une méthode pour consolider les candidats de relations produites précédemment en testant différentes approches.

# Références Bibliographiques :


[Amarger, 2015] Amarger F. Vers un système intelligent de capitalisation de connaissances pour l'agriculture durable : construction d'ontologies agricoles par transformation de sources existantes. *Thèse en Informatique de l'Université Paul Sabatier de Toulouse* soutenue le 18/12/2015

[AOS] Official FAO'sAgrontologyWebsite disponible à l'adresse suivante 31/08/2016: http://aims.fao.org/aos/agrontology.

[Ben Abbes, 2015] Ben Abbes S. Construction d'une cartographie de domaine à partir de ressources sémantiques hétérogènes. Thèse en Informatique de l'Université Paris 13 soutenue le 25/10/2013

[Bizer et al, 2009] Bizer C, Heath T, Berners-Lee T. "Linked data-the story so far". In Semantic Services, Interoperability and Web Applications: Emerging Concepts, 2009 pp 205--227.

[Fiorelli et al, 2014] Fiorelli M., Pazienza M. T., Stellato A., Turbati A. "CODA: Computer-aided ontology development architecture". In *IBM Journal of Research and Development*, Vol 58, N°2/3, pp 1--14, 2014.

[Jonquet et al, 2015] Jonquet C.,Dzalé-Yeumo E, Arnaud E, Larmande P. AgroPortal: a proposition for ontology-based services in the agronomic domain. *In les actes de IN-OVIVE'15:* 3ème atelier INtégration de sources/masses de données hétérogènes et Ontologies, dans le domaine des sciences du VIVant et de l'Environnement, 2015.

[Roussey et al., 2013] Roussey C., Chanet J-P, Cellier V., Amarger F. Agronomic Taxon. In *Proceedings of second international Workshop on Open Data (WOD 2013)*, 2013, BNF Paris

[Roussey et al., 2016] Roussey C., Bernard S., Pinet F., Reboud X., Cellier V.Gestion Sémantique des Bulletins de Santé du Végétal dans le projet Vespa. *In les Actes de l'atelier IN-OVIVE @IC* 2016.

[Szulman *et al*., 2009] Szulman S., Charlet J., Aussenac-Gilles N., Nazarenko A., Sardet E., and Teguiak H.V. . Dafoe: an ontology building platform from text or thesauri. In International Conference on Knowledge Engineering and Ontology Development (KEOD 2009), pages 1–4, 2009.